ED 426 587

DOCUMENT RESUME

ED 426 587 FL 024 983

TITLE	TELRI: Trans European Language Resources Infrastructure Newsletter, 1995-1997.
INSTITUTION	Trans European Language Resources Infrastructure.
SPONS AGENCY	Commission of the European Communities, Brussels (Belgium).
PUB DATE	1997-00-00
NOTE	215p.
AVAILABLE FROM	Web site: http://solaris3.ids-mannheim.de/telri/main.html
PUB TYPE	Collected Works - Serials (022)
JOURNAL CIT	TELRI: Trans European Language Resources Infrastructure Newsletter; n 1-7 Sep 1995-Oct 1997
EDRS PRICE	MF01/PC09 Plus Postage.
DESCRIPTORS	Computer Oriented Programs; Computer Software Development; Czech; Foreign Countries; Grammar; Information Technology; Language Maintenance; Language Planning; *Language Processing; *Language Research; *Languages; *Lexicography; *Machine Translation; Programming; Second Language Instruction; Second Languages; Uncommonly Taught Languages
IDENTIFIERS	*Copernicus; *Europe; Language Corpora

ABSTRACT

The first seven issues of the Trans European Language Resources Infrastructure (TELRI) newsletter, a publication of the COPERNICUS project funded by the Commission of the European Communities, date from September 1995 to October 1997. The first three issues contain articles in the origins of TELRI, its members, working groups, and events. TELRI's aim is to set up a network of leading national language and language technology centers in Europe. It brings together 22 institutions in 17 countries. Subsequent issues contain similar association-related information and articles on these topics: syntactic tagging techniques; the Czech National Corpus; issues in machine translation of Plato's "Republic" in a variety of Slavic languages; development of new lexicons; and development of multilingual technology. Some of the articles consist of summaries of conference papers on these topics. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 426 587

TELRI

TRANS EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE

Concerted Action in the Framework of the Copernicus Program

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Norbert
Vol 3

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Newsletter 1

September 1995

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL024983

BEST COPY AVAILABLE

Contents

1. Editorial (W. Teubert) ------------------ 3
2. The Copernicus Programme (P. Anderson) ----- 5
3. What does TELRI mean ? ---------------- 8
4. TELRI Partners ---------------------- 9
5. TELRI Working Groups ----------------- 21
6. TELRI Events ----------------------- 25

Coordinator:

Dr. Wolfgang Teubert
Institut für deutsche Sprache
P. O. Box 101621
D - 68016 Mannheim, Germany
phone: +49 62 58 428
fax: +49 62 58 415
e-mail: telri@ids-mannheim.de

Editors:

Prof. Eva Hajičová
Mgr. Barbora Hladká
Institute of Applied and Formal Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25
118 00 Prague 1, Czech Republic
tel.: +42 2 24 51 02 86
fax: +42 2 53 27 42
e-mail:
hajicova@ufal.mff.cuni.cz
hladka@ufal.mff.cuni.cz

Contributions to TELRI newsletter, and address corrections, should be sent to:

e-mail:
hladka@ufal.mff.cuni.cz
fax: +42 2 53 27 42

Editorial

Wolfgang Teubert, Coordinator of TELRI

Language engineering is the core of information technology, and information technology will be the key industry of the next decades. The information super highways conceived today will transport a variety of data, images, sounds, tables, figures, calculations, and process protocols. To make these data intelligible, they must be bound together by language. Without natural language processing information remains incomprehensible.

More than any other continent, Europe is multilingual. This situation provides a challenge to European language technology. We all want information to cross borders freely. But countries can only uphold their cultural and linguistic identity if all the relevant information is accessible and available in the national language(s). This is an important principle of the European Union today, and it also holds for all European nations. For the emergent European information society we have to develop a language technology that takes advantage of the multilingual challenge. It will have to support the production, revision, conversion, presentation, publication, documentation and last, but not least, translation of texts in technical and everyday language, and it will have to grant language-independent information retrieval by sophisticated interaction modes based on natural language. Language engineering in Europe will then play a leading role on the world market.

The quality of all language technology rests on the linguistic knowledge determining the algorithms of any natural language processing application. This linguistic knowledge is accessible in and by language resources. We find it in scientifically designed text corpora, in lexicons based on existing dictionaries and on corpus analysis, and we can extract it from textual and lexical resources and convert it into the form needed in application by powerful generic software, both language specific and language independent.

Language resources are the raw material of all language technology. The better they are the more expensive is their creation. The language industry, small and medium-sized enterprises in particular, often cannot afford to build them up. On the other hand, in all European countries there are focal language centres with a long tradition in the creation and application of language resources. What we need then, is a common infrastructure of (public domain) research and (private) industry. We need a common platform where providers and users of language resources come together, share expertise, discuss their needs, develop options and what is most important, exchange resources.

In some European countries such an infrastructure exists already, in others it is gradually being built up. But most of the work was (and still is) devoted to monolingual applications. There has not been much cross-border cooperation. This is why in Western Europe several efforts have been made to build up a common infrastructure that can serve the needs of multilingual language technology applications. In March 1995, the European Language Resources Association (ELRA) was set up with strong backing by the Commission of the European Community.

But Europe is larger than the European Union. Linguistic expertise, language resources and computational linguistics are highly developed in most Central and Eastern European countries. As well, we can observe here the emergence of a powerful, if still small, language industry. If we want to make Europe a competitor on the world market of language technology, we must build up a common infrastructure for the whole of Europe.

The Concerted Action TRANS-EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE (TELRI), a COPERNICUS project funded by the European Commission, brings together 22 institutions of 17 European countries, with strong links to relevant language centres all over Europe. These institutions pool their resources, build up multilingual expertise, develop generic tools for multilingual applications and create a strong permanent platform for successful cooperation between research and industry. TELRI was initiated in January 1995. Already, it has succeeded in setting up several multinational joint ventures with academic and industrial partners leading to concrete language technology products for today's and tomorrow's market. These projects will be presented in the next issues of this newsletter.

I strongly hope that our TELRI newsletter will make companies, organizations and institutes in research and industry aware of our activities. We need many partners with diverse backgrounds to develop a strong European network. Tell us about your needs. I am sure we will find a way to cooperate.

The Copernicus Program

by POUL ANDERSEN, DG XIII, European Commission

TELRI is one of the actions which were selected for funding after an open Call for Proposals, launched by the European Commission in 1994 under the COPERNICUS programme for Co-operation in Science and Technology with Central and Eastern European Countries.

Two types of actions were foreseen in the Call for Proposals :

- Concerted Actions
- Joint Research Projects

The European Commission received appr. 50 proposals in the area of Language Engineering, of which 6 proposals were for Concerted Actions.

The selection of proposals for funding was based on a reviewing procedure with external experts from both the European Union and Central and Eastern Europe. It was considered important to identify the best and most promising proposals in different sub-areas within Language Engineering, and to secure a reasonable involvement of all the eligible countries in Central and Eastern Europe.

The Call for Proposals was primarily targeted towards the following countries from Central and Eastern Europe: Albania, Bulgaria, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovak Republic and Slovenia.

TELRI is the only one of the actions in the area of Language Engineering which includes participants from all of these countries.

It was also possible for partners from the Newly Independent States of the former Soviet Union (NIS) to participate, and some of the other actions include a small number of participants from Belarus, Ukraine and Russia.

Besides TELRI, one more Concerted Action was selected for funding, ELSNET goes East - an extension of the existing ELSNET (European Network in Language and Speech) to Central and Eastern Europe.

Some of the selected Joint Research Projects involve many Eastern partners and resemble in this respect the Concerted Actions with a good potential for creating infrastructure and networking in more specific areas. This applies to

a Terminology project :

- ***PRACTEAST*** - Preparatory Actions for Terminological Assistance to Central and Eastern European Countries

- a Corpora project :
  MULTEXT-EAST - Multilingual Text Tools and Corpora for Central and Eastern European Languages.

- two Speech projects :
  BABEL - A Multi-Language Database.
  ONOMASTICA-Copernicus - Multi-language pronunciation dictionary of names in Central and Eastern European countries.

The other Joint Research projects have few Eastern partners and can be regarded as more 'concentrated' projects with a good potential for creating more immediately usable resources and/or applications :

- three projects concerned with Dictionary standards + coding
  CEGLEX - Central European GeneLEX model
  GRAMLEX
  BILEDITA - Bilingual electronic dictionaries and intelligent text alignment

- two projects within CALL (Computer Assisted Language Learning):
  BALTIC - Basic and advanced language transnational interactive course
  GLOSSER

- a Speech project :
  SQEL - Spoken Queries in European Languages

Some of the projects are in reality extensions of ongoing or recent (West) European projects to Central and Eastern Europe. This is most obviously the case for MULTEXT-EAST, ONOMASTICA-Copernicus, CEGLEX and as already mentioned for the Concerted Action ELSNET goes East.

More information can be found

- on World Wide Web, at
  http://www.fwi.uva.nl/fwi/research/vg2/illc/ege/cop.le.proj.html
  from where you will be guided on to local homepages for the individual projects and Concerted Actions.

- by contacting the project co-ordinators directly on the e-mail addresses which are listed below together with project reference numbers and short titles :

#200 ELSNET GOES EAST
Co-ordinator : Erik-Jan van der LINDEN, University of Amsterdam, Netherlands
e-mail : erikjan@fwi.uva.nl

#1202 TELRI
Co-ordinator : Wolfgang TEUBERT, Institut für deutsche Sprache, Mannheim, Germany
e-mail : telri@ids-mannheim.de

#58 ONOMASTICA - Copernicus
Co-ordinator : Mervyn JACK, University of Edinburgh, United Kingdom
e-mail : maj@ed.ac.uk

#106 MULTEXT-EAST
Co-ordinator : Jean VERONIS, University of Provence, Aix-en-Provence, France
e-mail : veronis@univ-aix.fr

#343 GLOSSER
Co-ordinator : John NERBONNE, University of Groningen, Netherlands
e-mail : nerbonne@let.rug.nl

#598 BALTIC
Co-ordinator : Albino BERTOLETTI, Giunti Multimedia srl., Milano, Italy
(no e-mail, phone +39-2-8393374, +39-2-8393408, fax +39-2-58103485)

#621 GRAMLEX
Co-ordinator : Eric LAPORTE, University of Marne-la-Vallee, France
e-mail : laporte@univ-mlv.fr

#787 PRACTEAST
Co-ordinator : Norbert KALFON, CL Servicios Linguisticos S.A., Madrid, Spain
e-mail : grupocl@ibm.net

#790 BILEDITA
Co-ordinator : Franz GUENTHNER, University of Munich, Germany
e-mail : gue@cis.uni-muenchen.de

#1032 CEGLEX
Co-ordinator : Antoine OGONOWSKI, GSI-ERLI, Paris, France
e-mail : Antoine.Ogonowski@erli.gsi.fr

#1304 BABEL
Co-ordinator : Peter ROACH, University of Reading, United Kingdom
e-mail : P.J.roach@reading.ac.uk

#1634 SQEL
Co-ordinator : Heinrich NIEMANN, University of Erlangen-Nirnberg, Germany
e-mail : niemann@informatik.uni-erlangen.de

What does TELRI mean ?

TELRI will set up a permanent network of leading national language and language technology centres in the whole of Europe. It will pool existing language resources, corpora, machine-readable dictionaries and lexicons, lexical databases, and generic software tools for the creation, re-use, maintenance, validation, and exploitation of linguistic data. It will complement these repositories with newly created multilingual resources, offering a wide range of language data to the NLP community. TELRI will establish a platform where research and industry meet, exchange resources and engage in product-oriented cooperation.

TELRI has a duration of three years (1995-1997). There are 22 participating institutions in 17 European countries (Albania, Germany, Great Britain, Slovakia, Italy, Bulgaria, the Czech Republic, Sweden, Slovenia, Romania, Estonia, France, the Netherlands, Latvia, Lithuania, Poland and Hungary). Links have been established with language centres elsewhere in Europe, with relevant European organizations and ventures, and with focal language institutions in other parts of the world.

TELRI is engaged in the following activities:

- Establishment of an Industrial User group representing software industry, publishers and translation services. TELRI partners and users carry out joint projects leading to marketable results.
- Documentation of language resources, generic software, institutions, projects and activities, to be made available on Internet.
- Validation and quality assessment of taggers, alignment software and homograph disambiguation software.
- Software design for language independent and language specific validation of language resources.
- Infrastructure awareness improvement by the TELRI newsletter.
- Joint presentation of service facilities.
- Organisation of a Seminar: Language Resources for Language Technology, as a dissemination and cooperation platform for research and industry.
- Creation of a special electronic TELRI network for online accessibility of all language resources among TELRI partners.
- Creation of a multilingual corpus and design of tools for the automatic detection of translation equivalents.

TELRI activities are organised in Working Groups of five to seven members.

TELRI Partners

- ***Institut für deutsche Sprache***
  Dr. Wolfgang Teubert
  Mannheim, Germany

The Institut für deutsche Sprache (IDS, Institute of German Language), founded in 1964, is the only research institution devoted to the documentation and description of German. It has a staff of more than 100 persons, including 65 academic researchers. It's projects explore grammar, vocabulary and pragmatic aspects of written and spoken contemporary German.

Research is based on corpora of different size, reflecting varieties of written and spoken language and well exceeding 50 million words. The IDS has developed powerful interactive corpus exploitation software for in-house and external on-line use. As the leading centre of German language resources, it is engaged in various European language technology projects and infrastructure activities.

- ***University of Birmingham, School of English***
  Prof. dr. John M. Sinclair
  Birmingham, Great Britain

Apart from the participation in TELRI, the Birmingham partners contribute to two international projects, for example EAGLES and PAROLE.

As part of the EAGLES project, Birmingham is reviewing current practices in the classification of texts in major European corpus projects. On the basis of these results, the aim is for a general classification scheme, proposing a text typology suitable for European corpus work. The normal kind of text typology deals with external criteria only, very little work has as yet been done on establishing or using internal criteria as a means for identifying text types. The work will make proposals for furthering the establishment of internal criteria for text typologies.

An important part of the PAROLE project is to establish links with interest groups on a national level. Birmingham has identified potential interest groups within the UK - both academic and industrial - and has met with them to discuss the possibility of collaboration and harmonization of language resources. Such groups include universities, publishing houses, government funding bodies.

A description of the PAROLE project, with links to the other partners, is available on the World Wide Web at http://clg1.bham.ac.uk/parole/, also at this address are details of a free tagging service for English texts. Birmingham is co-ordinator of the text typology subtask. The aim here is to produce specifications for the classification of texts based on the guidelines proposed by the EAGLES project. Birmingham is also co-ordinator of the subtask on spoken corpora. The aim of this report is to provide specifications for composition of spoken corpora, again based on the guidelines of the EAGLES report.

Both tasks will take into account the results of the NERC final report (which is now available from our ftp server - please contact parole@clg.bham.ac.uk for details).

■ *Hungarian Academy of Science, Research Institute for Linguistics*
Dr. Julia Pajzs
Budapest, Hungary

The Linguistics Institute of the Hungarian Academy of Sciences was founded in 1949. Its main goal is to support research in Hungarian linguistics and in general linguistics. The Institute has several departments, the most important research fields are: lexicography and lexicology, survey of spoken Hungarian, history of Hungarian language, Finno-Ugric linguistics, structural grammar of Hungarian, dialectal research, research in phonetics and phonology. One department of the Institute is a department of University Eotvos Lorand as well, the researchers of this department teach theoretical linguistics.

Research interests:

— compiling the dictionary of Hungarian
— collection from several sample texts which were carefully selected for this purpose by literary historians
— collection of spoken Hungarian several spoken text sample
— the task of revising the Concise Dictionary of Hungarian

■ ***Instituto di Linguistica Computazionale***
Prof. Antonio Zampolli
Pisa, Italy

Main institutional goals and mandate of the Institute headed by prof. Antonio Zampolli are the following:

— theoretical researches and applications in the area of computational linguistics and development of methods, techniques, instruments, basic linguistic researches, procedures.
— collaboration and technical-scientific assistance to institutions and organizations that use electronic elaboration of linguistic data for scientific and application-oriented goals.
— creation/distribution of resources and linguistic data and proposal/distribution of standards and generalized procedures.
— development of the relationship and co-operation with international organizations working in the same area and with centres of other countries.
— promotion of activities aiming at the diffusion of the scientific and technical knowledge in the area of computational linguistics.

■ ***Adam Mickiewicz University, School of English***
Prof. Jacek Fisiak
Poznan, Poland

The School of English was established in 1966 and is part of the Adam Mickiewicz University.

In the domain of computational linguistic analysis, the Institute works on English, Polish and, to a lesser extent, other European languages.

■ ***Warsaw University, Institute of Informatics***
Prof. Janusz Bień
Warsaw, Poland

This is a Computer Science Institute and a part of the mathematical faculty. The research on natural language processing was initiated at the Institute by Janusz S. Bień in late sixties; the research has been carried in a very close although usually informal cooperation with prominent Polish linguists, such as prof. Zygmunt Saloni and prof. Marek Swidziński. Stanisław Szpakowicz's formal desription of Polish syntax, developed in the Institute in 1978, is still widely used by other Polish teams, and his experimental parser was at that time one of the largest programs ever written in Prolog. Janusz Bień worked, among others, on taxonomical approach to Polish morphology; the review of his book (published in 1991) can be found in 'Journal of Slavic Linguistics' 2(2). Krzysztof Szafran's morphological analysis developed in 1993, thanks to the use of the most recent linguistic results, covers now not only over 120000 words from the great dictionary of Polish, but also new and potential words. At present Bień and Szafran are working on a new parser of Polish.

■ ***Romanian Academy, Center for Machine Learning,***
Natural Language Processing & Conceptual Modeling
Prof. Dr. Ioan Dan Tufis
Bucharest, Romania

The Center has been established in 1994 to conduct basic research in artificial intelligence and to promote international scientific research. It has a core of permanent staff (12 persons, 10 of them being experienced researchers), affiliates (7, out of which 5 are reputed scientists from abroad) and a variable number of temporary (contact-based) collaborators, mainly MSc or PhD students.

The main research projects of the Center are natural language processing, multistrategy learning, and computer-aided instruction. Some of the addressed research problems are:

— development of computational models of language
— development of a general method for multistrategy task-adaptive learning
— development of general methods for intelligent tutoring integrating AI technology with pedagogical fundamental principles

The main research areas being pursued at the Center in natural language processing are as follows:

— morphology and lexicon
— parsing in unification grammars
— logic and knowledge representation for language processing
— abductive language processing
— learning world knowledge by integrating different strategies

Collaborative Research: MAC-ELU, MAC-PAILab, FLUENT2, MULTEXT--EAST, ELSNET Goes EAST, ILP

■ ***Bulgarian Academy of Sciences, Institute of Bulgarian Language***
Prof. Dr. Iordan Penchev
Sofia, Bulgaria

The FOLG group at the Institute for Bulgarian language was set up in 1994 with the purpose of laying the foundations of the computational study of the Bulgarian language.

Current working projects of the group are:

— a large data base for the Bulgarian language (including an annotated corpus of texts of appr. 20 million words, a syntactico-semantic dictionary, a semantic analyzer and a morphological parser for Bulgarian).
— multilingual corpora, dictionaries and formal descriptions, oriented towards Computer Assisted Language Learning AND contrastive investigations.
— automatic spell-checkers and grammar-checkers.
— a corpus of 4 million words for the Bulgarian language.
— a trilingual translation corpus (Bulgarian, English, French) of 1 million words.
— a package of programmes for processing large monolingual and multilingual corpora.
— simple corpus analysis programmes (corpus statistics and ranging, concordances, a lemmatization environment).

■ ***Bulgarian Academy of Sciences, Center for Informatics and Computer Technology, Linguistic Modelling Laboratory (LML)***
Prof. Dr. Elena Paskaleva
Sofia

LML was set up in July 1987 in the framework of the Center of Infirmities and Computer Technology (CICT), Bulgarian Academy of Sciences. The Center's main objective was to coordinate scientific effort within the Academy and to promote international cooperation in the area of theoretical and practical problems of the new generations of computers.

LML Research Program

At LML research concerning NL is interdisciplinary-integrating linguists, logicians and computer scientists. Theoretical investigations concern:

— syntax and semantics of natural languages (semantics of tenses, verb frames, etc.);

- formalisms for knowledge representation (conceptual graphs, feature languages, nets and modal languages as unifying frameworks);
- logical foundations of reasoning with imperfect information;
- problems of database theory and object-oriented methodology.

Practical work is concentrated on:

- development of linguistic processors, based on different linguistic theories, and for pursuing different goals, e.g. full syntactic parsing or only partial one when concerned with a practical grammar checker;
- linguistic knowledge bases of different types and scope, e.g. computer dictionaries, morphological systems, etc.;
- intelligent information-retrieval and decision support systems based on AID principles, methodology and tools, in particular for aiding the translator of technical texts, or the practicing lawyer.

LML staff currently (spring 95) consists of 9 researchers. The main research activity of LML is the construction of the basic components of the Bulgarian Linguistic Knowledge Base envisaged for the (hopefully not so distant) future.

The creation of this linguistic knowledge base rests upon the following basic principles:

- aiming at exhaustive models of linguistic knowledge. In view of the obvious trade-off between completeness of knowledge and depth of the language level processed, the sequence of computer model realizations chosen is from the text via morphology and lexical data to syntax and semantics;
- attacking the deep language levels not through spectacular (albeit fragmentary) illustrative models, but through gradual accumulation of linguistic knowledge from real large text corpora by means of an intelligent user -oriented interface;
- attacking the surface language levels via complete computer models, e.g. a paradigmatic model of Bulgarian morphology with a very large LDB (with an envisaged scope of 60 000 entries); a system for processing and knowledge acquisition from large texts corpora obtained from desk-top publishing systems.

National Projects: LARGE BULGARIAN LDB, SUPERLINGUA, CONCEPTUAL GRAPHS, A LANGUAGE FOR DOMAIN KNOWLEDGE MODELLING, INFERENCE CONTROL INFORMATION IN THE KNOWLEDGE REPRESENTATION SYSTEMS

International Projects: TELRI, GLOSSER, BILEDITA, LATESLAV, ELSNET goes EAST

Institute for Dutch Lexicology
Prof. Piet Van Sterkenburg
Leiden, Holland

Dutch is the mother-tongue of ca. 21 million people in the Netherlands and Belgium. One of the fruits of cultural co-operation between North and South is the Institute for Dutch Lexicology INL at Leiden (the Netherlands). The INL, established in 1967, is an autonomous private foundation, financially supported by the governments of the Netherlands and Belgium. The INL has close relationships with State University Leiden. The director is Prof. dr. P.G.J. van Sterkenburg. Overall staff: 45 fte: lexicographers, electronic data processing (EDP) department, computerlinguists, auxiliary staff.

The general objectives of the INL are to construct linguistically annotated electronic text corpora, particularly in the Dutch language, and to compile linguistic products (dictionaries, lexical) on the basis of the text corpora, by use of computers.

Projects in progress at the INL are:

- Woordenboek der Nederlandse Taal WNT (Dictionary of the Dutch Language on Historical Principles WNT).
- Elektronisch WNT (Electronic WNT).
- Vroegmiddelnederlands Woordenboek VMNW (Dictionary of Early Middle Dutch), being compiled in an automized environment (lexicographical workbench).
- INL Language Database project. The electronic text corpora of present-day Dutch comprises over 150 million words. This language database will be developed towards an INL Integrated Language Database of 12th-21st Century Dutch, comprising linguistically annotated texts, lexicographical and linguistic data for the various centuries, linked in a linguistically senseful way.
- Dutch Spelling Guide.

The INL participates in the EC funded projects PAROLE (NERC), MECOLB, EAGLES, TELRI.

INaLF, CNRS
Dr. Pierre Lafon
Saint Cloud, France

INaLF is a research and service unit of Centre National de la Recherche Scientifique (CNRS), the public organization in charge of managing scientific research in France. The oldest component of INaLF was founded in 1960, with the initial task of writing a 16 volumes dictionary, the Tresor de la Langue Francaise (TLF), that would be to the French language what the Oxford English Dictionary is to Shakespear's mother tongue.

INaLF activities revolve around three main themes:

1. Text databases

INaLF is the creator of the FRANTEXT database which provides round the clock on-line access to a corpus of more than 180 million words covering 5 centuries of French literature, representing more than 3,000 complete texts from Renaissance to nowadays. On-line access has been provided (under certain conditions) to more than 80 libraries within France and to foreign institutions in Belgium, Finland, Portugal, Canada, Germany, Japan, and Sweden.

2. Creation of a lexical database for French

This database will group all the data of INaLF, present and future. All the pieces of information on French words will be gathered under a unifying header set. The core of this database will be the future computerized version of the Tresor de la Langue Francaise, around which will be integrated the by-products of the other projects undertaken at INaLF.

3. Development of historical lexicology

- writing of a dictionary for middle-age French (covering the period from 1350 to 1500), the building of a database for pre-classique French, the collecting of information for the study of the history of French vocabulary.

■ *Charles University, Faculty of Philosophy, Institute of the Czech National Corpus*

Prof. František Čermák
Prague, Czech Republic

In 1992, a group of linguists and mathematicians from various Czech Universities and from the Institute of the Czech Language, Academy of Sciences, Czech Republic, initiated the activities of the Computational Fund of the Czech Language. Among the main objectives of the Institute, established on the background of this initiative in 1994, there is the creation of a large general corpus that should become a versatile basis for all sorts of research and applications.

Charles University, Institute of Formal and Applied Linguistics
Prof. Eva Hajičová
Prague, Czech Republic

The Institute of Formal and Applied Linguistics was established in 1990.

The Institute collaborates very closely with the Institute of Computational and Theoretical Linguistics at the Faculty of Philosophy, Charles University (headed by Dr. Vladimír Petkevič). Both institutes continue in the work of the former group for computational linguistics founded by Petr Sgall and existing at Charles University since 1959.

Research interests, resources and expertise

— dependency syntax
— morphological analysis, POS tagging, syntactic tagging, algorithmical topic-focus identification
— morphological analyzers for Czech, English, lemmatization for full-text databases
— collection of daily newspaper texts
— procedure for automatic identification of topic and focus, explicit description of syntactic dependency relations
— English-Czech, Czech-Russian machine translation.

University of Goteborg, Department of Swedish
Dr. Martin Gellerstam
Goteborg, Sweden

The Department of Swedish is divided into for sections: Scandinavian languages, Swedish, Lexicology and Natural language processing. The section of LEXICOLOGY is directed towards theoretical research as well as lexicographic applications (we have published a comprehensive Swedish dictionary, Svensk ordbok (1986), a dictionary for immigrants and many other lexicographic works). The major part of our lexicographic work is based on corpora collected by the LANGUAGE BANK (founded in 1975), that builds up a large reference base of Swedish and makes data available for different kinds of users. Lexical data is integrated in our lexical database (GULD = Gothenburg University Lexical Database).

The research of our NLP section is directed towards the development of linguistic tools for analysing large corpora and making use of analysed linguistic material in different kinds of applications.

Comenius University, Computational Linguistics Laboratory,
Faculty of Education
Dr. Vladimír Benko
Bratislava, Slovakia

Comenius University in Bratislava is the oldest university in the Slovak Republic. It was founded in 1919 and follows the university tradition of the Academia Istropolitana which was established in Bratislava by Matthias Corvinus, the Hungarian King, in 1467.

The Faculty of Education prepares teachers for Basic Schools and in some subjects also for Secondary General Education Schools.

Computational Linguistics Laboratory, a small research unit founded in 1992, is aimed at carrying out NLP research and developing LI applications.

The main questions to be addressed are as follows:

1. creation of formal representations of language, computational interpretation of text
2. creation of universal tools and re-usable resources to support application development
3. application of research results in processing large text resources (quantitative text analysis, summarization), computer-aided language learning, creation of specialized lexicons, machine translation.
4. two projects being carried out in the framework of the Slovak Grant Agency for Science funding programme:
    1. Formal Models for Computerized Processing of Slovak Language (co-ordinator Eduard Kostolanský)
    2. Slovak Corpus and Lexical Database (co-ordinator Vladimír Benko).

Slovak Academy of Science, Ľudovít Štúr Linguistics Institute
Dr. Alexandra Jarošová
Bratislava, Slovakia

The scholarly activity of Ľudovít Štúr Linguistics Institute is concentrated on the basic research of the Slovak language.

Following projects are being carried out at present:

1. Slovak in Contacts (eventually in Conflicts) with other Languages: the interference of Slovak speakers in the contact regions, their verbal behaviour, national selfidentification, auto- and heterotypization.
2. Language/Speech Culture in the Theory and Practice: the functioning of the literary language in various forms of public communication, the rela-

tion between the literary language and other forms of national language, general and particular problems of terminology .

3. The System of Contemporary Slovak: functional stratification of Slovak language, the analyse of grammar/lexicon relations.
4. Slovak Corpus and Lexical Database.
5. Dictionary of the Contemporary Slovak Language.
6. Historical Dictionary of the Slovak Language.
7. Dictionary of the Slovak Dialects.
8. Atlas Elaboration of the Slavonic Languages in an International Co-operation.

"Jozef Stefan" Institute, Laboratory for Language and Speech Technologies
Tomaz Erjavec
Ljubljana, Slovenia

The Laboratory for Language and Speech Technologies was established in 1988 and is part of the "Jozef Stefan" Institute. The Laboratory is engaged primarily in research of Slovenian feature-based syntax and morphology, speech generation and (computational) logic.

Tartu University, Department of General Linguistics
Prof. Haldur Õim
Tartu, Estonia

Tartu Research Group In Computational Linguistics is an interdepartmental group where participate people from the Department of Estonian and the Department of Computer science. In the 80s the Group worked mainly in the frames of artificial intelligence, dealing with text understanding and dialogue modelling. In this context we also created some experimental programs of morphological and syntactic analysis of Estonian.

Since the end of 80s the work of the Group has centered on creating the computer corpus of contemporary Estonian, according to the classical principles of Brown and LOB corpora. We have put together 1 million words of texts, and at present the semi-automatic process of corpus tagging is on the way. A program of the morphological analysis of Estonian has been developed.

University of Latvia, Institute of Mathematics and Computer Science

Dr. Andrej Spektors
Riga, Latvia

The Institute of Mathematics and Computer Science was established in 1959 as Computing Centre of research character. It is the main institution in Latvia for basic research in computer science and mathematics.

The Institute has close co-operation in research, development, and advanced education with more than 10 research institutions in West Europe, the USA, and Japan.

The main research directions in the Computer Science division of IMCS are:

— computational learning theory and inductive program synthesis;
— complexity theory and probabilistic algorithms;
— specification languages;
— program testing automation;
— artificial intelligence and natural language processing.

Center of Computational Linguistics at Vytautas Magnus University

Dr. Ruta Marcinkeviciene
Kaunas, Lithuania

Not much could be said about our Center as it was founded only last year. Initiated by linguists it has an aim of complication of the first Lithuanian large spoken and written language corpora and promotion of corpus-based linguistics. There are four of us - two linguists and two informatics - all working part time in the Center. The first more tangible research results after the corpora are compiled should be the Lithuanian dictionary of Collocations and translation dictionaries based on it.

Much more should be said about our university bearing the name of grand duke Vytautas (1350-1430). Founded in 1922 it was the main university in the period between the two World Wars. Closed by the Soviets it was reopened only in 1989 and now is one of twelve higher schools of Lithuania. During the first five years it formed its structure of five faculties, comprising about 2000 students aiming at bachelor, master and doctor degrees and now is striving for unification of teaching and research, science and humanities.

Working Groups

WG1 TELRI USER GROUP

Co-ordinator: Wolfgang Teubert

TELRI is aimed at promoting cooperation between public domain research institutes and language industry. Particularly small and medium-sized enterprises, software developers, publishers or translation services, cannot afford to build up their own language resources. They have to rely on affordable resources provided by TELRI partners or similar institutions.

Only real projects aimed at the development of a marketable and concrete product will reveal which kinds of resources are needed. Therefore, each TELRI partner is asked to link with three companies either in a joint venture, or by supplying resources or other services to a commercial project. All TELRI partners take part in this effort, thus laying the foundation for a permanent platform where suppliers and users of language resources meet, share their knowledge, exchange data, and design new computational tools.

WG2 DOCUMENTATION

Co-ordinator: Ruta Marcinkeviciene

The aim of WG2 is to collect, document and disseminate relevent information on

1. language resources such as corpora, lexical databases, machine readable dictionaries plus lingware
2. language engineering organizations of all possible types: providers of data, users of data or of a mixed type
3. projects, meetings, conferences and other activities in the field of language engineering
4. relevant literature.

The outcome of our WG as well as all national TELRI partners' work should be a distributed knowledge database available both in an electronic and printed form. The following steps are to be taken in order to create the above mentioned database:

— the data to be documented defined and relevant actors identified,
— the requested data collected,
— the database designed and implemented,

— the database access mean implemented and partners as well as other countries informed.

WG3 NEWSLETTER

Co-ordinator: Eva Hajièová

The main task of the working group is to prepare and publish in regular intervals (three times per year) TELRI Newsletter informing the academic community, their industrial partners and also the prospective users about the activities of individual TELRI working groups, about available resources and about methods for their processing. The Newsletter helps in this way to make the communication between all interested parties easier and more effective.

WG4 TELRI SEMINARS

Co-ordinator: Julia Pajzs

The task of this working group is the organization of three "TELRI seminars". The first one, named "The European Seminar Language Resources for Language Technology" will take place in Tihany, Hungary 15-16 September, 1995. The aim of the seminars is to bring together linguistic software developers, companies concerned with multi-media electronic publishing, terminology and translation services, dictionary makers etc. Several joint venture research projects will be presented here and some invited speakers will give papers on regional scenarios from the global language technology market.

So far our working group succeded in collecting a mailing list for the first circular of the seminar, the location and time of the seminar was chosen. A hungarian small private service company was trusted by organizing the seminar. The first circular is to be distributed in a short time. The members of the seminar working group will meet in Hungary in July to decide on the final program and to solve any organizational problems.

WG5 LINGWARE ASSESSMENT

Co-ordinator: Tomaz Erjavec

Working Group will assess the performance of specified lingware under controlled conditions: corpus alignment software (1995), taggers (1996), and homograph disambiguation software (1997).

WG6 TELRI SERVICE POOL

Co-ordinator: Pierre Lafon

Working Group will pool existing service activities of TELRI partners and develop a design for a streamlined presentation of common service actvities, including issues as charges, contracts and copyright problems.

WG7 TELRI NETWORKING

Co-ordinator: Vladimír Benko

Working group will build up a dedicated electronic network (within INTERNET) between TELRI partners to enable exchange with and mutual access to each partner's language resources, and it will define standards for operation. Once in operation, this network will be open to TELRI User Group and Advisory Board members, and, by individual agreement, to other interested parties.

WG8 LINKING

Co-ordinator: Wolfgang Teubert

This Working Group pools and documents all relevant European and international links of all TELRI partners with institutions, organizations, and projects, and establishes new links.

Since some important language centres (in Croatia, Serbia and the Commonwealth of Independent States) are left out of the COPERNICUS programme this Working Group is used to have these centres participate in TELRI activities so they can become full members easily once a permanent infrastructure will be established.

Also, only seven European Union countries are represented in TELRI. It is necessary to establish links with the key centres in the other EU and European Economic Area countries to develop a common European infrastructure. For this purpose, TELRI will work closely with PAROLE; other important links that are being established are with EAGLES, ELSNET and ELRA.

Because in Western Europe COPERNICUS activities are still felt to be only marginally important when compared to other EU activities, this Working Group must increase the EU language resources community's awareness of the available assets in Central and Eastern Europe. One way used to increase awareness and visibility of TELRI is to create links with important centres overseas, to organize joint activities with them and to organize a world-wide platform for the exchange of resources, expertise and the design of language technology applications.

■ *WG9 ORGANIZING JOINT RESEARCH*

Co-ordinator: John Sinclair

The group will run through very modest exercises of a research nature in order to learn how to work together and share resources.

One of the interesting areas for the group is that of parallel corpora, where each corpus is the translation of another, and alignment can be achieved at the sentence level. We decided to choose a classical text, on the grounds

1. that translations into most of the project languages would be likely to exist
2. that no language would be privileged by being the original.

We chose Plato's Republic. Members of the group are currently trying to find a suitable translation. If it is not in electronic form already (as the English translation by Jowett is), funds will be sought from the Co-ordinator for the cost of scanning or rekeying. Around the mid-year, we will review progress.

In the second phase of the project, the "tools" group intends to work on alignment software, so we hope to be able to take advantage of that work.

■ *WG10 USER NEEDS*

Co-ordinator: Andrej Spektors

The work of WG10 is aimed at the establishing of the user needs, identifying of the existing resources, preparing of the proposals for a repository of lexical resources (such as corpora, machine readable dictionaries and lexical data bases) and the developing of lingware.

Every member of the group works within his country with the potential users he chose. Afterwards the summarizing of the results will be made and the common features will be generalized.

■ *WG11 PERMANENT INFRASTRUCTURE*

Co-ordinator: Antonio Zampolli

Working Group will prepare a proposal for the design of a permanent infrastructure to extend and complement the infrastructure that will be set up for the European Union language industry.

TELRI Events

when? where? what?

- *JANUARY 13 &14:*
  TELRI Inaugural Meeting
  Mannheim / Germany

Plenary Meeting, Steering Committee meeting and meetings of TELRI working groups

- *FEBRUARY 28:*
  Meeting of the ELRA (European Language Resources Association) Interim Steering Committee
  Luxembourg / Luxembourg

Wolfgang Teubert takes part in the meeting. As a member of this committee he furthers the interests of the Central and Eastern European countries.

- *MARCH 27:*
  Meeting of the ELRA Interim Steering Committee
  Brussels / Belgium

Wolfgang Teubert takes part in this meeting.

- *APRIL 3-16:*

Visit to China by Wolfgang Teubert as co-ordinator of TELRI Working Group "Linking" to form contacts with promising language centres in China (Institute of Applied Linguistics, part of the Chinese State Language Commission in Beijing: Prof. Feng Zhiwei / Institute of Natural Language Processing of the Jiao Tong University, Shangai: Prof. Yang Huizhong / Departament of Computer Science of the Tongji University, Shangai: Prof. Sheng Huanye and Prof. Wu Quidi). Results are conceptions of some low-scalejoint projects that could lead to medium-term cooperation between TELRI members and Chinese partners.

- *MAY 15:*
  Meeting of the ELRA Steering Committee
  Brussels / Belgium

Wolfgang Teubert takes part in the meeting.

- *MAY 18-21:*

Visit to Belgrade by Wolfgang Teubert as co-ordinator of Working Group "Linking". The visit to Belgrade led to concrete plans for the participation of the Chair of Computer Science of the Faculty of Mathematics at Belgrade University, Prof. Dusko Vitas. The goal is - even though Serbia, Croatia, Bosnia and Macedonia are excluded from the COPERNICUS programme - to use existing contacts to link key institutions there to the TELRI Concerted Action.

- *MAY 27:*
  TELRI Steering Committee Meeting
  Florence / Italy

- *MAY 27:*
  Meeting of TELRI Working Group 2 "Documentation"
  Florence / Italy

- *MAY 27:*
  Meeting of TELRI Working Group 8 "Permanent Infrastructure"
  Florence / Italy

- *JUNE 10:*
  Meeting of TELRI Working Group 7 "Networking"
  Bratislava / Slovakia

On the second weekend of June, the 2nd TELRI WG7 meeting (originally scheduled for April) took place in Bratislava, Slovakia. All the WG7 member sites were represented, mostly, by the "instead-of's": Oliver Jakobs (BIR), Cyril Belica (MAN), prof. Janusz Bień (WAR), and Dr. Eduard Kostolanský and Vladimír Benko (BRA1). Tomaz Erjavec (LJU2), the WG5 coordinator, was also present at the meeting.

The discussed topics covered experience with electronic communication among the TELRI partners in general, the current state of the facilities provided (TELRI mailing list established in Mannheim, TELRI WWW home page currently maintained by Tomaz Erjavec in Ljubljana and Edinburgh), specific problems of "slow" TELRI sites and the question of using national character sets in e-mail communication and WWW access. It has been agreed that the TELRI WWW home page is to be moved to Mannheim where it will be maintained and updated on a regular basis. The "Internet How-To Manual" for TELRI partners that is being prepared will concentrate mostly on basic procedures and suitable "entry points" where additional information is to be found.

- *JUNE 23 & 24:*
  Meeting of TELRI Working Group 3 "Newsletter"
  Prague / Czech Republic

- *SEPTEMBER 11-14:*
  TELRI Plenary Meeting, Steering Committee meeting, meetings of all TELRI Working Groups
  Tihany / Hungary

- *SEPTEMBER 15 & 16:*
  TELRI Seminar
  Tihany / Hungary

WHAT IS TELRI

The Concerted Action TRANS-EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE (TELRI), is a COPERNICUS project funded by the European Commission. TELRI has a duration of three years (1995-1997). It brings together 22 institutions of 17 European countries (Albania, Germany, Great Britain, Slovakia, Italy, Bulgaria, the Czech Republic, Sweden, Slovenia, Romania, Estonia, France, the Netherlands, Latvia, Lithuania, Poland and Hungary).

TELRI will set up a permanent network of leading national language and language technology centres in the whole of Europe. It will pool existing language resources, corpora, machine-readable dictionaries and lexicons, lexical databases, and generic software tools for the creation, re-use, maintenance, validation, and exploitation of linguistic data. It will complement these repositories with newly created multilingual resources, offering a wide range of language data to the NLP community. TELRI will establish a platform where research and industry meet, exchange resources and engage in product-oriented cooperation.

Links have been established with language centres elsewhere in Europe, with relevant European organizations and ventures, and with focal language institutions in other parts of the world.

TELRI's WWW Document

Detailed information about TELRI and its activities is available through the World Wide Web (WWW) at the following URL:

http://www.ids-mannheim.de/telri/telri.html

FOR INFORMATION:

Inquiries about TELRI may be addressed to:

Dr. Wolfgang Teubert
Institut für deutsche Sprache
P.O. Box: 101621
68016 Mannheim, Germany
Phone: +49 621 1581 437
Fax: +49 621 1581 415
email: telri@ids-mannheim.de

TELRI

NEWSLETTER
No. 2

Issue No. 1 | Issue No. 3 | Issue No. 4 | Issue No. 5

Contents

- **Editorial** (W. Teubert)
- **TELRI Partners**
- **News from TELRI Working Groups**
- **TELRI Events: Hungary, Tihany seminar**

| TELRI Partners | News from TELRI Working Groups | TELRI Events

Editorial

Wolfgang Teubert, Coordinator of TELRI

One year ago TELRI was set up by 22 institutions in seventeen European countries. Meanwhile, more institutions have joined, if only as members of the Advisory Board: Belgrade, Zagreb, and, quite recently, Moscow. TELRI is a Concerted Action. Its primary goal is to pool existing language resources, corpora, and lexicons and to make them available to the growing NLP community. New resources are being created; all resources will be standardised. Together they will allow for the development of a new generation of powerful multilingual language technology applications. They are already used for corpus-based dictionaries and lexical databases. The TELRI Resources Catalogue is available on the WWW (http://www.ids-mannheim.de/telri/telri.html). The success of the TELRI network depends on the organisation of national networks bringing together academic research and commercial language industry. Therefore, TELRI and the partner project ELSNET Goes East joined forces to conduct an indepth survey on the leading actors in the field of language technology, their resources, and their activities. Data collection is about to begin. Results will be made quickly available on Internet and in printed form.
Our first European Seminar, "Language Resources for Language Technology", conducted in Tihany, Hungary on September 15 and 16 clearly demonstrated the need for a continuous platform for research and industry. All applications more sophisticated than spelling checkers heavily rely on linguistic data and knowledge extracted from the data. Successful applications are those that have an accuracy rate of more than 95%. The higher your goal on this scale, the more linguistic knowledge required. This knowledge is available at academic research centres. Thus, if you want to upgrade your language technology products

available at academic research centres. Thus, if you want to upgrade your language technology products contact us now.
Meanwhile, we have selected some areas of ongoing research. We set out to produce electronic text versions of Plato's "Republic" in as many languages as possible. Nine (including Chinese) are already available. This small, but very helpful parallel corpus, will be used to test corpus alignment software and to find new ways to detect translation equivalents of multi-word units, collocations, and phraseologisms. Another area is the validation of textual and lexical resources: here we want to complement similar activities in Western Europe and to assess generic tools such as alignment software. Validation of language resources will have to be based upon linguistic specifications like morphosyntactic tagsets for corpora or data categories for lexicons. Some of these are language independent (e.g., parts of speech), others (e.g., gender) are language specific. TELRI will contribute to a joint European register of linguistic specifications to be used as a standard for language resources and generic software. Whoever would like to hear more about these activities or would like to participate in them, please contact us.

| Editorial | News from TELRI Working Groups | TELRI Events

Telri Partners

In the first issue of our newsletter we published information about TELRI partners. For technical reasons , the following descriptions could not be included there, so that we publish it in the present issue.

Institute of Linguistics and Literature
Prof. Dr. Bahri Beci
Tirana, Albania

The Institute of Linguistics and Literature was founded in 1972. The first nucleus of the Institute was the section of Language, literature and folklore in the Institute of Studies (1946). With the extension of the scientific activity there were set up five sectors: grammar and dialectology, lexicology and lexicography, terminology, history and literature and folklore.
The fundamental task of the Institute of Language and Literature in the field of linguistics is the study of Albanian language and its dialects in their actual state and on the historical plan, especially the study of the national literary language, elaboration of scientific grammar, compiling of explanatory dictionaries of Albanian language, elaboration of terminology and compiling of multilingual dictionaries of terminology, treatment of onomastic problems, and culture of language.

Centre of Scientific Research of the Slovene Academy of Arts and Sciences
Institute of the Slovene Language "Fran Ramovs"
Primoz Jakopin
Ljubljana, Slovenia

Under the cover project "Lexicology, grammar and dialectology of the Slovene language," the Institute is currently engaged in research on the following projects:

(1) the word corpus of the contemporary standard language (the one-volume dictionary of the contemporary standard Slovene language; the orthographical dictionary; the backwards dictionary of the

contemporary standard Slovene language, based on the five-volume dictionary of the con temporary standard Slovene language; dictionary of standing expressions of the contemporary standard Slovene language; synonymity in the contemporary standard Slovene language; the transfer of the word- list Besedisce slovenskega knjiznega jezika to machine-readable form; the preparation of the five-volume dictionary of the contemporary standard Slovene language for edition in electronic form)

(2) the morphological and word-formational analysis of the contemporary standard Slovene language

(3) comparative and etymological investigations of the Slovene language (the publication of volume 3, and the preparation of volume 4, of the Etymological Dictionary of the Slovene Language

(4) historical dictionaries (the dictionary of the writings of Slovene protestant writers of the 16th century; the dictionary of the one-time standard language based on the Prekmursko dialect)

(5) Dialect atlases and dictionaries (the preparation of dialect maps for the Slovene linguistic atlas; the participation in the international projects The Pan-Slavic Linguistic Atlas [OLA] and The European Linguistic Atlas [ALE]. The preparation of dialect dictionaries: of the Kostelsko dialect, of the dialect of Zadrecka dolina between Gornji grad and Nazarje. Monographs: the tonemes in the word-formation of the contemporary standard Slovene language, contrasted with the dialect of Vnanje Gorice; the dialect of Kropa)

(6) Terminological dictionaries (the terminological dictionary of law; the dictionary of general technical terminology; the terminological dictionary of medicine; the terminological dictionary of veterinary sciences; the terminological dictionary of railways).

The cover project "Lexicology, grammar and dialectology of the Slovene language" is a long-term research project designed to ensure (a) the full inventorisation of the Slovene lexical material, and (b) a systematic analysis plus interpretation of the language facts at all levels of grammar, from a historical as well as from a descriptive point of view. The main purpose of the cover project is to produce the basic publications in the field of the Slovene language, which publications (1) will deepen the insight into the Slovene language as it is now and as it was in the past, (2) contribute towards the equitable treatment of the Slovene language in international linguistic circles.

Editorial | TELRI Partners | TELRI Events

News from TELRI Working Groups

WG1 TELRI USER GROUP
Co-ordinator: Wolfgang Teubert

In this Working Group, we had a very ambitious goal: each TELRI partner was to engage in three small-size joint ventures with small- and medium-sized language industry enterprises. In general, our partner was to contribute the necessary language resources and the linguistic knowledge required while the company would produce the result: a NLP application, a dictionary, or some other product that could be marketed. We wanted to show that cooperation between research and industry is not only possible but that it also can be profitable for both partners involved.

Our goal could not be reached everywhere. In some countries, commercial language industry is still in its infancy. The small software houses have very little money to invest, and they are looking for quick returns. However, good and solid language technology applications need a longer breath, and some companies will still have to learn that you need more than a traditional small dictionary and a smart programmer to develop sophisticated NLP software.
National networks will bring together research and industry, providers and users of language resources. Their experience, knowledge, and information will be shared; however, we also need national programmes to encourage this kind of cooperation and to induce language industry to engage in more sophisticated applications. In a number of successful projects, our TELRI Seminar has demonstrated that cooperation between research and industry is not only possible but also profitable.

WG2 DOCUMENTATION

Co-ordinator: Ruta Marcinkeviciene

The most important piece of news about WG2 is that it has started coordinating information collection and documentation with the other two projects - ELSNET and ELSNET goes EAST, involved in the same activities. Representatives of all the three projects met in London, August 12 and decided to pool their efforts in creating a widely accessible database of the language and speech technology groups in industry and academia. They discussed what sort of information still has to be gathesed, where information is available and how to task should be approached in a practical way. They also agreed to develop a common set of questionaires. In adition, in order to avoid duplication of activities, to increase the response rate and to use EC funds more effectively it was decided to assign the Western European countries to ELSNET and to divide up the Central and Eastern countries between ELSNET goes EAST and TELRI.

WG5 TOOL AVAILABILITY

Co-ordinator: Tomaz Erjavec

The initial work plan for WG5 (Tool assessment; 1995 - tagger assessment) was found to be unattainable, due to the lack of human, language and computational resources. At the Tihany meeting it was therefore decided to change the name and tasks of the WG: *WG5 Tool Availability*.

WG5 will aim to increase the availability of language engineering tools by:
_ making available, via WWW;
_ providing, via WWW, the public tools of WG5 members and TELRI partners;
_ improving such tools by adapting them to various languages and platforms.
The WG will concentrate on multilingual or language independent tools for developing and exploiting textual resources.

The above goals will be pursued in cooperation with:
_ WG9 Joint Research (`Cascade' project);
_ WG7 Networking (organisation and utilisation of WWW);
_ WG10 User Needs (tools for corpora validation);
_ Copernicus JP MULTEXT-East.

Current results:
_ a WWW page providing information on various public and commercial language engineering tools (http://nl.ijs.si/~tomaz/pub_tools);
_ Gothenburg site has started efforts to make their lexicon building tool robust and publicly available.

Plan:
_ WG5 Ljubljana site, in cooperation with Prague will make the MULTEXT tools publicly available and work on adapting these tools for various languages and platforms;
_ WG5 Prague site has agreed to make their morphological analyser publicly available;
_ WG10 plans to use the MULTEXT tools for assessing corpora;
_ WG9 Birmingham site has plans to make available their suite of corpus handling tools (`Cascade').

WG6 SERVICE POOL

Co-ordinator: Pierre Lafon

Due to specific circumstances, the coordination of this Working Group was temporarily attributed to Dan Tufis. Under a short notice he prepared the meeting in Tihany and afterwards this report on the decisions and agreements that have been made concerning future work.
The main issues came into three categories:
_ types of services the TELRI consortium might offer
_ legal aspects concerning services
_ future plans

A. TYPES OF SERVICES THE TELRI CONSORTIUM MIGHT OFFER

This issue was raised mainly considering the already established associations with similar aims. It was agreed that given the fact that TELRI includes (either as partners or associates) representatives from most of European Countries creates a very strong advantage in acting as a bridge to Central and Eastern European language Market for the associations that are almost exclusively based on Western Europe and overseas countries. Besides the standard services ensured by such associations as ELRA or LDC some others were discussed:
_ evaluation of language resources (including assesment of lingware) for own-language
_ porting/extending software to cover missing features
_ implementation teams - design teams
_ linguistic assistance
In a previous phase, IDS-Manheim compiled a list of existing resources at the sites of the TELRI partners, which it was decided to be updated regularly and further extended with a precise statement of the services (including rates, fees, copyright problems etc) that could be ensured by the holders of these resources.

B. LEGAL ASPECTS CONCERNING THE SERVICES

This issue was quite a hot one as the legislation (mainly the copyright law) appears to be quite different in the member countries of TELRI consortium (in Romania, there is no copyright law yet). The difference in the legislations might be a source of difficulties in ensuring a unitary service. This is particularly true of corpora-related services, containing material subject to the copyright regulations. A specific user-agreement (used by INL for the use of an retrieval system) was discussed trying to point out the important items that might be included into a generic TELRI-service agreement. A general advise: avoid being too specific. According to the most existing copyright regulations, for each use not agreed on in writing, permission has to be requested. The main ideas (contributed by T. Kruyt, based on the INL experience) are the following:

KEY ELEMENTS
_ parties involved in the agreement: - names of provider and the user

_ topic of the agreement
_ purpose of the agreement
_ permission
_ time schedule agreement
_ details of delivery
_ guarantees

C. FUTURE PLANS

The members of this WG with support form all the other members of TELRI will gloss over existing and potential services which could be of interest both internally and externally the project. This information should be made available to an as large audience as possible. The WG6 should act as a broker for the pooled services. It was decided to have one of the Newsletters dedicated to the services offered by the TELRI consortium. The special issue of the TELRI Newsletter will present also the available linguistic resources. By the end of November a questionnaire would be circulated among the TELRI partners for updating the information already collected and for extending it with new types of services.

WG8 Linking
Co-ordinator: Wolfgang Teubert

This Working Group aims at broadening TELRI as a European platform for the creation, enrichment, distribution and exchange of high quality monolingual and multilingual language resources. Working Group Linking is setting up close operational ties with the recently established PAROLE Consortium, an association of all leading national language centres in the European Union. This Consortium has compiled linguistic specifications for generic written resources and is now creating comparable reference corpora and lexicons. TELRI will engage in these activities on a complementary basis. TELRI also has formed relations with Professor Vladislav Mitrofanovich Andrjuscenko and his Institute for Russian Language. It is the Russian centre of corpus linguistics. Furthermore it has established links with ELSNET and ELSNET Goes East. Together these projects will conduct an indepth survey of actors, data and activities in the field of language resources in Central and Eastern Europe (including the New Independent States).

WG10 USER NEEDS
Co-ordinator: Andrej Spektors

The results of the work on questioning potential users and their needs were discussed. During the exchange of the results of distributed questionaires we found some features common for Central and Eastern European countries. The existing resources are mostly in text file format, and they have not been validated. Potential users just start to show their interest in linguistic software . User needs are very close to the results achieved by NERC and PAROLE projects . The only problem in the CEE countries is that user needs are not yet strictly formulated. A plan for the future work has been accepted. The plan is based on the development of validation methods. The possibility to validate corpora according to SGML, TEI and EAGLES standards and recommendations is proposed as the main objective.

Editorial | TELRI Partners | News from TELRI Working Groups

TELRI Events

TELRI Seminar, Tihany, Hungary, September 15-16 1995

The European Seminar "Language Resources for Language Technology" was the first of its kind organized by TELRI. It took place in the Institute for Limnology in Tihany, Hungary, 15-16 September 1995.
The aim of the seminar was to bring together scholars, software - lingware developers and end-users to exchange information. Several state of the art papers were presented by the invited speakers, the covered fields ranged from speech processing through machine translation to corpus application. We even had a presentation on the language resources and softwares in China, and we could also broden our knowledge on the American market for linguistic data. Some talks dealt with other language engineering COPERNICUS project. The representatives of the Hungarian government and the European Comission also gave lectures.
The members of TELRI user group presented some joint venture case studies, among them we saw a modern railway dictionary, two spell-checkers, a new project for collecting neologisms from corpora, and the use of the computer fund of Russian. We also had a chance to see several demonstrations, some of them presented by TELRI members or partners, some of them by external participants. One of the most challenging was the demonstartion of the LANGMaster Multimedia System for Language Teaching. The more then 70 participants came from 24 countries. Beside TELRI members and partners we had guests from universities, research institutes, private software companies and publishing houses. Beyond the usual advantage of conferences, the seminar offered a possibility of communication between software developers and users, and as a result, some business agreement were settled.
The seminar was a succesfull attempt to establish a new forum for researchers in the field of corpus linguistics and natural language processing, and for the possible users of their results.
Julia Pajzs

DEMONSTRATIONS

Demonstrations of NLP systems of most different kinds were one of the most interesting parts of the Tihany Workshop Programme. We publish de scriptions of some the demonstrated systems for information, where we are, and for inspirations.

Primoz Jakopin (Ljubljana)
EVA _ A TEXTUAL DATA PROCESSING TOOL

A text editing program, which has, from 1985 on, evolved into a tool, which served for processing of a sizeable number of textual corpora and preparation of dictionaries in the Slovenian academic environment, is presented. EVA started on a Sinclair Spectrum (EVE), has been ported to ATARI ST machine in 1986 (STEVE); DOS version is in use since 1991. Porting to Windows NT/Windows 95 is under way.
EVA has been designed, from the start on, to be as flexible as possible, to allow the accomodation to different needs and situations by the user himself. It is more or less self contained, with its own keyboard, screen characters, DTP mode, graphics editor and an OCR facility. To conform to modern character set standards such as UNICODE EVA has a capability to process either 8- or 16-bit characters. If a line of text contains only characters with codes below 256, it is, in RAM as well as on disk stored as 8-bit; if, on the other hand, it contains one or more characters with codes above 255, it is stored as a 16-bit entity. All internal line and data record buffering is of course 16-b it. Data base functions include general purpose

routines such sorting or searching and more specialized function such as splitting of text into sentences, wordwise translation and markup or computation of entropy.
Currently EVA is also used in production of a lemmatization dictionary of Slovenian, based on the 93.500 entries long Dictionary of the Slovenian Literary Language. So far nouns (54.522 lemmas to generate 468.281 word forms) and adjectives (22.861 lemmas and 277.831 words forms) have been completed.

Prof. Dr. Elena Paskaleva (lingware) / Bojaka Zaharieva (software)(Sofia):
THE ARDUOUS TAGGING OF HIGHLY INFLECTIVE LANGUAGES
(NON-ENGLISH; NON-LATIN ALPHABET; NO MORPHO AUTOMATION)
The object of the demonstration is the system SUPERLINGUA. SUPERLINGUA is a tagging tool for highly inflected languages in extreme conditions: if the morphological component is missing or unusable or if the language is NLP virgin (not having been processed at all). The system is language independ ent, the tagging is flexible and friendly and a special interface is provided for the optimal distribution between the system's and user's linguistic knowledge. The programming language is CLIPPER 5.2 in DOS environnment. The system is supposed to be available in public domain in 6-9 months in DOS and WINDOWS environnment. The product is made entirely by the software specia lists of the Laboratory of Linguistic Modelling.

Dr. Andrejs Spektors (Riga):
MULTILINGUAL OFFICE-TERM DICTIONARY

Purpose: The system was designed as a tool for people who have problems with writing documents in foreign languages (Latvian, Russian, English).
How it works: The system is developed for PC computers and works under DOS. It consists of three items: general term dictionary, thematic dictionary and case generation tool.
Developer: AI Lab. of Institute of Mathematics and Computer Science, University of Latvia.
Can be obtained: by agreement with AI Lab

Dr. Andrejs Spektors (Riga):
AUTOMATED CASE GENERATION SYSTEM FOR LATVIAN

Purpose: The system is used to generate wordforms in Latvian for different purposes, e.g., morphological analyzer, vocabulary for spelling-checker, computer aided language learning. During the work a database of anomalous words was developed.
How it works: The system is developed for PC computers and works under DOS. It includes case generation for nouns, adjectives and verbs. The system can be used in different modes, i.e., demonstration or learning mode and wordform generation mode for lexicon.
Developer: AI Lab. of Institute of Mathematics and Computer Science, University of Latvia.
Can be obtained: by agreement with AI Lab

Dr. Andrejs Spektors (Riga):
MODEL OF LATVIAN MORPHOLOGICAL ANALYSER
AND REDUCTION TO THE BASE FORM

Purpose: The system is developed for further usage in syntactic analysis and for lexicalization
*How it works:*The system is developed for PC computers and works under DOS. It analyses sentences

separately and returns base forms of each word as well as part of speech and grammatical information. For homonyms all possible solutions are produced.
Developer: AI Lab. of Institute of Mathematics and Computer Science, University of Latvia.
Can be obtained: by agreement with AI Lab

Dr. Andrejs Spektors (Riga):

ELECTRONICALLY TRACTABLE LATVIAN DICTIONARY

Purpose: providing user interface to monoligual Latvian_Latvian dictionary.
How it works: The dictionary works under MS Windows. The list of all words in the dictionary is presented to the user and can be scrolled or incrementally searched for some word. In another window area dictionary entries can be viewed, and the user can easily get possible base form(s) of any word form present in dictionary entries.
Developers: New Mexico State University, U.S.A. (prof. J.Reinfelds), Univerity of Latvia, Department of Baltic languages, AI Lab. of Institute of Mathematics and Computer Science, University of Latvia.
Can be obtained: by agreement with NMSU

Jan Laciga (ByllBase, Prague):

BYLLBASE - A FULL-TEXT RETRIEVAL METHOD USING LINGUISTIC METHOD

There are principally two approaches to the task of information retrieval of textual data: (i) to select the text according to indexes (key words) assigned to each text, or (ii) to retrieve a word or combination of words directly in the texts and thus to select documents where the issues referred to by the given (string of) words are discussed.
The system developed by our company belongs to the type (ii), which we consider to be more convenient for large scale applications. We had to develop a system specifically designed for Czech because the systems available mostly for English are not applicable: the inflectional character of Czech (in contrast to English) brings problems connected with the rich abundance of forms of a single lexical item.
The first commerically available system for text retrieval for Czech , called ByllBase, has been developed in cooperation with the group of computational linguistics at Charles University in Prague and its special feature is an integration of the lemmatizer of Czech into the system. This lemmatizer makes it possible also to distinguish among homonyms. This enables the user to formulate the queries in a natural form, it speeds up the whole process and lowers the requirements on memory capacity for the auxiliary files. At the present stage, we make amendments to the semantic analysis to make it possible (without a human interference) to distinguish among homonyms.
ByllBase is used nowadays at such big institutions as the Czech saving bank Èeská spoøitelna, the Czech National Bank, the city council of Brno, Bratislava, some industrial plants, editorial offices etc.
One of the sucessful installation of ByllBase is the legal system ASPI, a most complex and widespread automatic retrieval system of legal documents in the Czech Republic and in Slovakia, which contains Czech legal documents and legal literature since 1811.
The system was developed by ByllBase in close cooperation with the researches of the team of computational lunguistics at Charles University, Prague.

Prof. Dr. Dan Tufis (Bucharest):

UNIFICATION-BASED IMPLEMENTATAION OF A WIDE COVERAGE ROMANIAN MORPHOLOGY

The MAC-ELU is an integrated unification-based system aimed at developing reversible linguistic

descriptions. It consists of a morphological analyser/generator, a chart parser, a head-driven generator and a transfer module, all of them relying on unification mechanisms for dealing with grammatical constraints. The morphological processor works on a continuation-classes basis, with the usual clustering of morphemes into distinct dictionaries (called continuation classes). Successful transitions from one cluster to the other, corresponding either to analysis of a word-form structure or to the generation of a word-form by concatenation, are constrained by specific restrictions introduced by means of a powerful macro-definitions mechanisms. The implementation of Romanian morphology (and NP analysis/generation) will be exemplified and the structure of the lexical entries will be discussed. Further development plans will be mentioned.
The ELU system was developed by ISSCO (Rod Johnson, Mike Rosner, Graham Russel, Afzal Balim, Amy Winarske) , running in ALLEGRO-COMMON LISP on SUN machines. The MAC-ELU version of it, was ported in Machintosh Common Lisp running on Macintoshes by Dan Tufis and Octav Popescu. The ported version includes some new facilities, a menu based interface and it was code-optimized in order to ensure a reasonable response time for the smaller machines. The Romanian morphology was implemented by Dan Tufis, Octav Popescu, Lidia Diaconu, Calin Diaconu and Ana-Maria Barbu. Most of the NP rule set is due to Lidia Diaconu, Calin Diaconu and Cristian Dumitrescu.
For information on how to obtain this system, contact: Dan Tufis: e_mail address: tufis@u1.ici.ro

Prof. Dr. Dan Tufis (Bucharest):
GULiveR: A GENERALIZED UNIFICATION LR PARSER

This is an extension of Tomita's GLR parser, with the significant departure from the original algorithm of working with feature-based grammars. Also, the data structuring introduced by Tomita to take care of the conflicting entries in the LR-tables (graph-structured stack, packed shared parse forests) have been enhanced to allow for non-monadic grammar categories (DAGs). To deal with the complexities problems raised by unification extension of the GLR parser we used special data structures (virtual copying vectors). Although test data are not available yet, we expect GUL̂iveR to be provide very good response time in real applications.
This parser was developed in 1992 by Dan Tufis and Octav Popescu. The initial implementation (Golden Common Lisp for PC) was ported this year on Macintoshes (MCL2.0.1) by Stefan Bruda and Mihai Ciocoiu.

Prof. Dr. Dan Tufis (Bucharest):
PARSING PORTABLE LABORATORY

This demo, based on a larger education software, called PAIL, is a nice tutorial system for learning about parsing. It presents two different paradigms: the old ATN procedural approach and the declarative one supported by a parametrized chart-parser. The graphical interface, the steppers, the browsers, animated graphic, demos and on-line documentation made this system a very effective educational tool, highly appreciated by students. The larger PAIL system (which includes besides NLP modules, several other interesting Artificial Intelligence systems - theorem proving, rule-based systems, neural networks, back propagation, constraint satisfaction programming, inductive learning, genetic algorithms) was initially implemented at IDSIA-Lugano by Rod Johnson, Mike Rosner, Paolo Cattaneo and Fabio Baj (with contributions from some others) in Allegro Common Lisp on SUN workstations. The system was ported in MCL for Macintoshes by a joint team consisting of Mike Rosner, Paolo Cattaneo from IDSIA and Dan Tufis, Octav Popescu, Stefan Trausan and Adrian Boangiu from Romanian Academy and ICI.

Prof. Dr. Dan Tufis (Bucharest):
KRIL - A KNOWLEDGE REPRESENTATION INTERFACE TO AN INTERLINGUAL NATURAL LANGUAGE GENERATOR

This system is based on a natural language generator (ALLP), developed by Sue Felshin and Stuart Malone of MIT ATHENA's group. ALLP takes as input a highly verbose interlingual representation of a syntactic structure (GB flavoured) and produces natural language text. KRIL allows for a highly conceptually specified input. On the basis of the linguistic information already present in the lexicon it automatically generates the lengthy structures needed by ALLP. The overhead added by KRIL is less than 10% of the overal generation time (the medium response time is below 1 second for a 6-8 word sentence). The KRIL generator makes the linguistic processing fully transparent to a client application (such as, for instance, an intelligent tutoring system in second language learning).The KRIL interface was implemented by Dan Tufis.

Jan Prùcha (Dr. Lang Group, Prague):
THE LANG MASTER TEACHING SYSTEM

LANG Master Teaching System is a system of computer programs and instructions designed for the teaching of foreign language by means of a computer. The whole project consists of three main points:

1. LANG Master Technology
LANG Master Technology is a procedure that prepares data for LANG Master Presentation. Explicitly it converts a chosen language course or dicitionary from a book form into the form of computer data.

2. LANG Master Presentation
LANG Master presentation is a powerful multimedia application designed to present LANG Master computer courses and dictionaries for the teaching of foreign language.

3. RE-WISE Method
The aim of the method is to keep in the student's memory all the expressions learnt and, at the same time, to minimalise the frequency of revision.

Vladimír Benko (Bratislava):
Concise Dictionary of the Slovak Language _ Electronic Version

The Concise Dictionary of the Slovak Language (KSSJ Krátký slovník slovenského jazyka, VEDA, Bratislava 1987) is a one-volume everyday-use explanatory dictionary of present-day Slovak, covering some 55,000 headwords (36,000 main entries). The last `paper' edition appeared in 1979.
The electronic version of KSSJ is based on the typesetting tape of the dictionary's first edition that had been transformed into the (slightly) tagged MRD form, (extensively) validated and (manually) updated to match the second printed edition. The reformated data have been indexed by means of the WordCruncher corpus processing package.
The current MRD version of KSSJ has been used as one of the reference sources to compile the new Slovak Synonyms Dictionary (Synonymický slovník slovenèiny, VEDA, Bratislava, in print), as a tool for various research projects in Slovak lexicology and as teaching material at the Comenius University's Faculty of Education. The CD-ROM version of KSSJ is on consideration to appear simultaneously with the third edition of KSSJ, that is being prepared to appear in the end of 1996.

Dr. Truus Kruyt (Leiden):

ACCESS TO A LINGUISTICALLY ANNOTATED 27 MILL WORD CORPUS OF DUTCH NEWSPAPER TEXTS VIA INTERNET

The Institute of Dutch Lexicology INL is a research institute subsidized by the Dutch and Belgian governments. Corpus development at the INL dates from the mid-seventies. Up to 1990, the INL text corpora were mainly developed for lexicographical purposes. Presently, they are used for a broad variety of research and applications. INL text corpora of present-day Dutch include two linguistically annotated corpora which can be consulted via Internet: the 5 Million Words Corpus 1994, which covers a variety of topics and text types, and the 27 Million Words Newspaper Corpus 1995. The retrieval program developed for the latter will be demonstrated.

Characteristics of the 27 Million Words Newspaper Corpus 1995:
The newspaper texts, dating from 1994 and 1995, were obtained in machine-readable form, on a contract basis with the publishing company. The contract specifies the conditions of use. The texts were input for automatic linguistic encoding. Part of speech (POS) and headword were automatically assigned to the word forms in the electronic texts by a lemmatizer/POS-tagger developed by the INL. Most of the data has not been corrected, neither on the level of the proper text, nor on the level of POS and headword. The linguistically encoded texts were loaded into an on-line retrieval system developed by the INL. Queries may concern the whole corpus, or a subcorpus defined by the user along the parameters year and month of publication. The system allows the user to search for single words or word patterns, including some, still rather primitive, predefined syntactic patterns which can be revised by the user. Search definitions may include references to word forms, POS and head words, both separately and in combination by use of Boolean operators and proximity searches. Output data most often is a list of items, or a series of concordances with a user-defined context size. With limitations due to copyright, the output of searches can be transferred to the user's computer by e-mail (it is not allowed to transfer complete texts or substantial text fragments). Among the other facilities are the use of wild cards and various sorting facilities.

Access to the 27 Million Words Newspaper Corpus 1995:
Consultation of the corpus is free for non-commercial purposes. Please contact the director of the INL, Prof. dr. P.G.J. van Sterkenburg, about the conditions for commercial applications. To get access to the corpus, an individual user agreement has to be signed. An electronic user agreement form can be obtained from our mailserver Mailserv@Rulxho.Leidenuniv.NL. Type in the body of your e-mail message: SEND [27MLN95]AGREEMNT.USE. Please make a hard copy of the agreement form, sign it, keep a copy yourself, and return a signed copy to: Institute for Dutch Lexicology INL, P.O. Box 9515, 2300 RA Leiden. After receipt of the signed user agreement, you will be informed about your username and password. Use of a VT 220 (or higher) terminal, or an appropriate terminal-emulator (e.g. Kermit) is recommended. If you need additional information, please send an e-mail message to Helpdesk@Rulxho.Leidenuniv.NL, or send a fax to Mrs. dr. J.G. Kruyt (31 71 27 2115).

JOINT VENTURE STUDIES

One of the aims of the TELRI project is to promote cooperation between academia and industry. Contributions devoted to some joint ventures that were presented at the Tihany seminar have shown the usefulness of such cooperation.

Dr. Truus Kruyt (Leiden):
A NEW DUTCH SPELLING GUIDE

Dr. Truus Kruyt and Prof. Dr. Sterkenburg,
Institute for Dutch Lexicology INL, Leiden, The Netherlands.

Dutch Spelling Guides: 1954, 1990, 1995

The most recent official Dutch spelling guide, compiled in order of the governments of the Netherlands and Belgium, dates from 1954. The Belgian Spelling Resolution of 1946 and the Dutch Spelling Law of 1947 were applied to the Dutch and Flemish vocabulary by a Dutch-Belgian spelling committee consisting of 12 experts in the field.

In the past decades, this spelling was considered too complicated. New spelling principles were proposed by several official and inofficial committees, without any success up to October 1994, when the Dutch and Belgian governments agreed on not too radically changing principles for a spelling revision. A new guide is being compiled in order of the Dutch-Belgian government body `Nederlandse Taalunie' by the Institute for Dutch Lexicology INL, and will be published in printed and in electronic form by the `Staats Drukkerij en Uitgevery' SDU.

In the meantime, in 1990, the INL and the SDU published an unofficial spelling guide, including the ca. 65.000 entries of the 1954 guide and additionally ca. 30.000 new entries, which for the most part represent words that have come into use since 1954. INL was responsible for the contents of the guide, SDU for its publication. The division of the revenues is established by contract.

Dutch Spelling Guides 1990, 1995 and Language Resources

The spelling guides not only list entries with their correct orthography, but also provide information on spelling variants, hyphenation, genus, conjugation and inflexion, etc. Both the selection of entries (macrostructure) and the contents of the information categories per entry (microstructure) are determined by evidence coming from a collection of electronic written language resources, containing over 150 million words, available at INL. The resources include three text corpora (5, 27 and 50 million words, resp.) which are linguistically annotated for headword and part of speech (POS) and accessible on these parameters by a retrieval program (cf. demo '27 Million Words Corpus of Dutch Newspaper Texts via Internet'). The word forms in the additional textual resources needed still to be lemmatized and the texts to be made accessible for the purpose. Main criteria for the empirical basis of the information in the guides are frequency and coverage.

INL acquires the textual materials from several publishing houses on a contract basis. Due to the use of different systems for text preparation by the publishing houses, the acquired texts have different formats. The texts were to be converted, filtered for information not relevant for this application, and formally harmonized to some extent, so as to make them appropriate as input for further processing and consultation.

Future cooperation

Apart from this one, the INL resources have proven to be of interest for other product development projects of commercial companies. Future cooperation could be supported and improved by more uniform standards, at the levels of text preparation, data exchange and consultation of linguistic data.

Gabor Proszeky (Morphologic, Budapest):
HUMOR, a Morphological System for Corpus Analysis

Humor, a reversible, string-based, unification approach for lemmatizing and disambiguation has been introduced for both corpus analysis in the Research Institute for Linguistics, and creating a variety of other lingware applications, like spell-checking, hyphenation, etc. for the wide public. The system is language independent, that is, it allows multilingual applications: besides agglutinative languages (e.g. Hungarian, Turkish) and highly inflectional languages (e.g. Polish, Rumanian) it has been applied to languages of major economic and demographic significance (e.g. English, German, French).

The basic strategy of Humor is inherently suited to parallel execution. Search in the main dictionary, secondary dictionaries and affix dictionaries can happen simultaneously. What is more, in the near future it is going to be extended by a disambiguator based on the same strategy. This is a new parallel processing method of various levels (higher than morphology) called HumorESK (Humor Enhanced with Syntactic Knowledge). Both Humor and HumorESK have a very simple and clear strategy based on surface-only analyses, no transformations are used; all the complexity of the systems are hidden in the graphs describing morpho-syntactic behavior.

Humor is rigorously tested by "real" end-users. The Hungarian version has been used in every-day work since 1991 both by lexicographers and other researchers of the Research Institute of Linguistics of the Hungarian Academy of Sciences, and users of word-processing tools (Humor-based linguistic modules have been licensed by Microsoft, Lotus, Inso and other software developers). The lemmatizer shares some of the extra features of Helyes, the speller derived from Humor, because lexicographers need a fault-tolerant lemmatizer that is able to overcome simple orthographic errors and frequent mis-typings. It is useful in analyzing Hungarian texts from the 19th century when the Hungarian orthography was not standardized.

Humor's Hungarian version the largest and most precise implementation contains nearly 100.000 stems which cover all (approx. 70.000) lexemes of the Concise Explanatory Dictionary of the Hungarian Language. Suffix dictionaries contain all the inflectional suffixes and the productive derivational morphemes of present-day Hungarian. With the help of these dictionaries Humor is able to analyze and/or generate several billions(!) of different well-formed Hungarian word-forms. The whole software package is written in standard C using C++ like objects. It runs on any platform where C compiler can be found.

Primoz Jakopin (Ljubljana):

RAIL-LEX SLOVENIA - A MODERN RAILWAY DICTIONARY

The two partners involved are Slovenske zeleznice, the Slovenian Railway (Railway Traffic Institute) and the Institute for Slovenian Language at the Scientific Research Centre of the Slovenian Academy of Sciences and Arts. The work on the project, Dictionary of the Railway Terminology (Zelezniski terminoloski slovar) began in January, 1994 and is to be completed by the end of 1998.

The dictionary is a part of a larger European undertaking, Rail-lex Europe, under way by coordinated efforts of 29 members of the UIC, Union internationale des chemins de fer (International Union of Railways). UIC consists of 97 railway and other transport organizations from Europe and other parts of the world. The aim of the Rail-lex project, which has so far, in 1994, produced an 11-language CD ROM Rail Lexic with over 12.000 keywords (English, German, French, Italian, Spanish, Esperanto, Hungarian, Dutch, Polish, Portuguese, Swedish), is to put together a modern, multilingual communications infrastructure, to promote links between railways themselves and between railways and the Industry, research and commerce and to contribute to the standardization of railway terminology. Rail-lex is coordinated by UICs European Rail Research Institute (ERRI), based in the Netherlands.

On Slovenian side head of the project is mag. Peter Verlic, leader of the team at the Railway Traffic Institute in Ljubljana, aided by Marjan Vrabl, who is leading the team in Maribor, the second largest Slovenian city, where a new set of railway codes, manuals and other documentation is being prepared. After Slovenia has become independent in 1991, the changes, needed to bring the railway closer to UICs standards, have to be made. The bulk of the keywords from Rail Lexic have now been translated, and together with additional keywords, which reflect the social and other specific circumstances in Slovenian railway they now form the first draft of a 15.000-keyword dictionary. It will be open to criticism from the railway staff and wider audience till end of 1996, when a revision from the side of the Institute for the Slovenian language will also be completed.

Norbert Volz (Institut für deutsche Sprache, Mannheim,
E-Mail: volz@mx300c.ids-mannheim.de):

CORDON _ CORPUS-ORIENTED DETECTION OF NEOLOGISMS

CORDON , a multinational concerted project jointly carried out by academical and industrial partners, aims to provide a modular, language-independent client/server software solution for the automatic detection of neologisms _ new words or multi-word-units denoting new concepts _ in texts using monitor corpora.

New concepts reflecting changes in culture, society, industry and science quickly show their influence to language. New words or multi-word-units emerge, enabling the integration of these concepts in the communication progress. The identification and documentation of those changes therefore is of major importance for maintaining the actuality of language resources, language processing tools and terminology databases.

Monitor corpora can be used to recognise and trace the changing patterns of collocations and similar phenomena that give clues to the emergence of new terms. Basically, two types of tools are needed for this purpose:

_ a tool to correlate lexical and terminological items with temporal intervals, based on frequency and distribution over text types; using statistical methods such as c^2-tests to assess the significance of noticeable irregularities in the distribution of words of a corpus within a certain time

_ a statistics-driven tool to establish context patterns for lexical and terminological items, reflecting their various usages, e.g. by the examination of the verbal environment of repeating instances of words, looking for repetitions and regularities within the environment.

A combination of these tools working on monitor corpora will enable the identification of "candidates" for neologisms, which then can be listed and processed for further analyses and applications.

The envisaged software product will be a minimal assumption, generic modular solution that any users can adapt to their own texts and corpora regardless of language. Possible applications will mainly be within lingware products, e.g. machine translation systems, multilingual termbanks, databases etc. *CORDON* will also prove useful for the automatic updating and expansion of natural language lexicons and translation memories.

The project consortium consists of four academic and four industrial partners. The academical partners will provide research facilities and staff. The industrial partners will be responsible for project management, supervision, validation, evaluation and assessment of the final product in order to guarantee maximum response to user needs.

Project duration will be two years. At the end of this phase, the result of the *CORDON* project will be a demonstrable robust prototype that will work on existing application and corpora.

The proposal for this project will be handed in under the current TELEMATICS call within the 4th Framework Programme of the European Commission.

Elena Paskaleva (Sofia):

EUROPEAN LANGUAGE RESOURCES AND THE COMPUTERIZED RUSSIAN LANGUAGE FUND

CEU RSS (Central European University-Research Support Scheme) has sponsored a project with 5 participants from 3 countries - 2 from CRLF, 2 from GMS-Berlin and 1 from LML (Linguistic Modeling Laboratory - Bulgarian Academy of Sciences). Limited resources have been granted for the application of

10 000 dictionary entries from Ozhegov's Dictionary of the Russian Language to the Russian part of the data in a METAL-type system for Machine Translation.

Top of this issue TELRI Main Page IDS Main Page

TELRI

NEWSLETTER
No. 3

Issue No. 1 | Issue No. 2 | Issue No. 4 | Issue No. 5

Contents

- Editorial
- Topic of this issue: Syntactic Tagging
- News form TELRI Working Groups
- Awareness Day in Bucharest
- TELRI Events - Birmingham Workshop

Syntactic Tagging | News form TELRI Working Groups | Awareness Day in Bucharest | Birmingham Workshop

Editorial

Wolfgang Teubert, Coordinator of TELRI

1. General Problems

In spite of the relatively smooth progress that TELRI made in the reported period of the second half year (7/95 - 12/95), our experience was that some overall conditions were not entirely beneficial to the objectives of this Concerted Action. We can identify the following problem areas:

- The constitution of Concerted Actions
- Coordination of related COPERNICUS activities
- Lack of support by Western European links
- The emergent EU infrastructure

1.1 The constitution of Concerted Actions

In the COPERNICUS Programme, the goal of Concerted Actions is to bring together as partners focal institutions in central and Eastern Europe (CEE) with their counterparts in Western Europe. Therefore, TELRI consists of 22 institutions in 17 countries, among them 12 Central and Eastern European countries. By membership in the TELRI Advisory Board, now three and soon some more institutes in more CEE and NIS (Newly Independent States) are closely linked to TELRI. As a typical infrastructure organization, TELRI does not carry out research, but tries to set up a common platform for the exchange of expertise, software, language resources, information, and new ideas, and to establish a common identity that will facilitate cooperation in projects aimed at multilingual language technology.

At the same time, all TELRI partners are involved in a number of research projects, within their own institution and on a multilateral basis. The projects provide funding for the actual research, and this creates a situation with which TELRI cannot really compete. Most of the TELRI budget goes into coordination and travel expenses. With few exceptions, reimbursement for work dedicated to TELRI is not possible.

Considering this situation, the high level of motivation, of efforts and of concrete results achieved in TELRI is quite remarkable. To keep this spirit alive, however, we will have to bridge the gap between research and infrastructure activities by preparing a small list of more research oriented projects that can complement TELRI's wider goals. The VALIDATOR proposal sub mitted to the last COPERNICUS call is a good example. In the second year of TELRI, we will, therefore, explore the possibilities for such multilateral projects which would have a beneficial impact on infrastructure while at the same time provide adequate funding for research work.

1.2. Coordination of COPERNICUS activities

In the area of language resources and language engineering, we find in the COPERNICUS Programme these action lines:

Concerted Actions for the creation of a pan-European infrastructure, Projects leading to concrete results (resources, tools, or applications), Awareness Days, workshops, summer schools, etc.

All these activities have their own individual profile. Still everyone involved agrees that even though efforts for concentration between certain activities already exist, more could and should be done. Perhaps a project complementary to UNGLINK in Western Europe could promote additional synergy effects.

1.3. Lack of support by Western European links

The first year of TELRI gave rise to the impression that some institutions, activities, and circles in Western Europe are not very fond of the idea of a pan-European infrastructure and not very supportive with respect to the needs of their counterparts in Central and Eastern Europe. TELRI is working hard to make the vision of pan-European cooperation attractive. But more encouragement by the European Commission is needed to open up existing Western European standardization, information, and distribution networks not just to select individuals in CEE but to each eligible institution of a fair basis.

1.4. The emergent EU infrastructure

The year 1995 saw the rapid growth of an operational infrastructure for language resources in the EU. ELRA was founded with substantial financial support by the European Community; the academic partners in the PAROLE I Project founded the PAROLE Association; and EAGLES delivered standards and specifications which are intended to be used in the whole of Europe. Some partners in CEE and NIS countries feel uneasy about these developments which will have an impact on them, but which they cannot join. More about this problem in the following chapter.

2. A Changing Environment for Language Resources

In the year 1995, we witnessed an explosion of data, images, sounds, tables, figures, process protocols, options, and visions distributed globally via ever-expanding information superhighways. If these data are to be intelligible, if they are to make sense, they must be bound together by language. Without natural language processing, information remains incomprehensible. For the emergent global information society, we have to develop a language technology that meets the multilingual challenge. It will have to support the production, revision, conversion, presentation, publication, documentation, and last, but not least, translation of texts in technical and every day language; and it will have to grant language-independent retrieval by sophisticated interaction modes based on natural languages.

Europe is determined to remain a multicultural and multilingual society. Where other information technology markets, like North America or Southeast Asia, can restrict themselves to monolingual or, at the most, bilingual applications, Europe has to develop a language technology that creates a truly multilingual information society by helping its citizens to overcome language borders.

Multilingual textual and lexical resources employing the same standards and closely linked in their composition are essential for the development of multilingual applications. Therefore, with the financial support of the European Commission, important steps for a language resources infrastructure in Western Europe were taken in 1995. On the organizational level, ELRA (European Language Resources Association) and, for written resources, the PAROLE Association were founded. The new PAROLE II project was prepared and finally accepted. ELRA delivered first recommendation for standards as well as guidelines and specifications. This had consequences on four levels:

Infrastructure level: Ties between focal language resources institutions were strengthened (PAROLE Association); links between academic research and private industry were established (ELRA).

Standardization and validation level: standards and specifications for text representation, lexicon markup, and morphosyntactic and syntactic features were adopted (EAGLES, but also MULTEXT, MECOLB, and PAROLE I); first outlines for the validation of written resources were designed.

Distribution level: ELRA was set up as a European distribution center for language resources.

Production level: PAROLE II was prepared with the goal of creating a first generation of comparable resources for multilingual applications.

These developments, which so far have largely left out CEE and NIS countries, demand adequate responses by TELRI. It was necessary to strengthen TELRI activities in the area of standardization, validation, and distribution, and to find ways to participate in the creation of comparable language resources. The following chapter will deal with these accommodations.

3. Accommodation of the TELRI Workplan

3.1. Introduction of new work items

The Working Group User Needs (Coordinator: Andrejs Spektors, Riga) completed its first survey on industrial user needs by late fall 1995. A final survey on user needs will be carried out by Working Group User Groups (Coordinator: Wolfgang Teubert, Mannheim) in 1997 on the basis of an analysis of joint ventures carried out by TELRI members. The Working Group of Andrejs Spektors has now adopted the work item Validation. First task is the preparation of a proposal for a COPERNICUS project VALIDATOR. This project will ensure participation of CEE and NIS partners in the design of written resources validation and, thus, establish a uniform and homogenous approach to validation in Europe.

The Working Group Seminars (Coordinator: Julia Pajzs, Budapest) was given the new work item Morphosyntactic Features. The organization of the Tihany Seminar made it clear that subsequent seminars will be organized by the local partner and Mannheim alone, thus, making a Working Group Seminars superfluous. On the other hand, the various (and rather heterogeneous) recommendations categories put forward by ELRA, MULTEXT, MECOLB, and PAROLE were not seen by TELRI partners to suit the peculiarities of Slavic and Baltic languages nor of Hungarian or Estonian. The reconstituted Working Group Morphosyntactic Features will endeavor to unify existing recommendations, complement them with necessary features of languages not yet covered, and propose a synthesis permitting various levels of granularity for different applications. It will seek to establish close links with all relevant activities mentioned as well as with COPERNICUS activities like MULTEXT East. This is a necessary step that has to be taken for the development of truly comparable pan-European language resources.

TELRI plans for 1996 to prepare a proposal for a new COPERNICUS project PAROLE East with the goal to create complementary standardized resources for such CEE and NIS countries where some resources which can be converted already exist. TELRI will set up an informal Working Group Bridge Dictionaries for a concerted preparation of localized versions of the COBUILD Student Dictionary (English entry words, but descriptions in local languages based on the English original). These dictionaries can, as electronic versions. easily be linked and, thus, converted into a multilingual lexicon. In addition, these individual projects will provide useful experiences for joint ventures between academic research and private industry; and it will also generate some income for participating institutions.

Working Group Documentation (Coordinator: Ruta Marcinkeviciene, Kaunas) is cooperating with ELSNET Goes East in the preparation of a new and comprehensive edition of a survey of CEE and NIS institutions, enterprises, and organizations active in the field of language resources and language engineering. TELRI will, in 1996, explore the feasibility of a project Multilingual Terminological Database for Language Resources and Language Engineering with partners all over the world.

3.2. Additional activities

The papers given at the Tihany European Seminar Language Resources for Language Technology will be published as a book.

Working Group Organizing Joint Research (Coordinator: John Sinclair, Birmingham) will link up with Working Group Multilingual Lexicons of the PAROLE Association with the goal of developing a methodology for the realization of translation equivalents. TELRI will expand and regularly update its Web pages and, in addition, set up and open TELRI list for increased visibility of the TELRI Concerted Action.

Editorial | News form TELRI Working Groups | Awareness Day in Bucharest | Birmingham Workshop

Topic of this issue: Syntactic Tagging

What Linguists May Expect And Require From Syntactic Tagging

Petr Sgall
Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics
Charles University,
Prague, Czech Republic

Under the given technical conditions, text corpora often are conceived of as containing not only data on part of speech appurtenance of the individual lexical occurrences, but also information on their morphemic values and syntactic functions. Only if this information is sufficiently rich and reliable, the corpus can serve as a valuable source for large scale exploitation in most different areas of research on language, including not only language and its structure itself, but also the theory of literature and neighboring disciplines.

One of the important questions is how to select and organize the grammatical data in the tagged corpus. It goes without saying that data on morphemics should be maximally detailed and that they should be patterned in such a way that it would be easy to use them with different theoretical frameworks. The latter issue is more complex in what concerns syntax and its relationships to semantics and pragmatics. In these domains, a theory independent approach to tagging cannot be understood as using only concepts independent on any linguistic theory, but rather in the sense mentioned, i.e. as applying sets of categories (with their values) and decision procedures that would allow the linguist using the corpus to classify the tag symbols in accordance with the needs of as many existing (or reasonably imaginable) theoretical frameworks as possible.

Another condition requires the tagging procedure to be simple and modular enough to make a semi-automatic approach possible. To this aim, the basic and most frequent phenomena should be described by means of relatively perspicuous categories and values, not too distant from an intuitive view of the sentence structure and of the grammatical properties of lexical units. From this it follows that errors occurring in the output of a first version of the tagging procedure (which contains a parser, perhaps based on a combination of grammatical and statistical steps) may be identified by individual checking and the quality of the procedure could be amended by solutions avoiding the most frequent errors.

A rather general assumption on which syntactic tagging may be based is that the syntactic relations (and several aspects of morphological information) in the prototypical case are expressed by morphs (prepositions and other function words, endings, or affixes), whereas surface word order serves to the expression of the topic-focus articulation. Also in English, French or Chinese the "given" (contextually bound) information usually precedes the "new" part of the contents of a sentence. The grammatical function of the SVO order certainly will be used in parsing languages in which this kind of configurational structure is present; however, it would not be appropriate to base the identification of syntactic relations on such a starting point also in cases where the word order is "free", i.e. not grammaticalized, be it in languages with a higher degree of "free" word order, or e.g. in the order of some adverbials in English.

Taking this assumption into account, we come to the following conclusions:

(i) the function words should be rendered in the system of tags by symbols indicating the corresponding functions, i.e. morphological values of the corresponding autosemantic words (e.g. values of tense, number, definiteness, degrees of comparison) and syntactic relations (specifiers, arguments and adjuncts, or complements, modifications, and so on); if possible, not only the differences between subject, direct, indirect and "second" object (the latter present e.g. in *Fred was elected the chairman*) are to be distinguished, but also several tens of kinds of adjuncts (adverbials such as Locative, Manner, Means, Condition, several Directionals, Temporal adverbials, and so on - corresponding to primary and secondary meanings of prepositions, subordinating conjunctions and other means); these values should be indicated in any case, be they expressed by function words, affixes, stem alternations or word order;

(ii) for every autosemantic occurrence other than the main verb of the sentence it should be indicated whether it is a complementation (argument, adjunct, etc., see above) of a certain head or a part of a coordinated construction (for which again the head it depends on would be specified);

(iii) the surface word order of the autosemantic lexical occurrences in the output of the tagging should not differ from their surface order; this would allow for an analysis of the topic-focus articulation of the sentence; if it is probable that the intonation center of the sentence (when read aloud) would be placed elsewhere than at the end of the sentence (as e.g., in English, with sentences containing such a word like *yesterday* after the verb and its complement, or with short sentences containing a cleft construction), the bearer of the intonation center (constituting the focus proper of the sentence) should be marked by a specific index.

Points (i) and (ii) ensure, at least to a certain degree, that for theories requiring a further classification of syntactic relations it will be possible to specify the additional specification (e.g. the subject of an active verb may be identified as an Actor, corresponding in a cognitive layer either to Agentive, or to Experiencer, Theme, and so on, according to the context; or it may be classified as the NP constituting an immediate constituent of the sentence).

The output of the tagging procedure may have the form of a bracketted string with indices (with every dependent word and every coordinated construction being enclosed in its pair of parentheses, an index of this pair identifying the syntactic function of the word or the kind of coordination, and a set of indices at each word indicating its morphemic values). Only in the exceptional cases in which the condition of projectivity (adjacency, continuity of constituents) is not met it would be necessary to indicate the position of the head e.g. by its serial number (this concerns especially the long distance dependencies).

Certainly, most parsers available today or in the near future will not go that far (e.g. in what concerns the oppositions of different functions of prepositions, or the identification of the intonation center). However, tagged corpora will make it possible to analyze the relevant syntactic issues in monographs, dissertations, etc., for individual languages and their groups, and we may hope that results of such research can then be used to amend the analytic procedures.

Formal Representation of Language Structures

Jan Hajic*, Eva Hajicová*, Alexandr Rosen**

**Institute of Formal and Applied Linguistics*
Faculty of Mathematics and Physics

***Institute of Theoretical and Computational Linguistics*
Faculty of Philosophy
Charles University
Prague, Czech Republic

Abstract

Building treebanks is a prerequisite for various experiments and research tasks in the area of NLP. Under a recently awarded grant,1 we are developing (i) a formal definition of a (dependency based) tree, and (ii) a mid-size treebank based on this definition. The annotated corpus is designed to have three layers: morphosyntactic (linear) tagging, syntactic dependency annotation, and the tectogrammatical annotation. The project is being carried out jointly at the authors' Institutes.

1 The Current State and Motivation

Recent decades have seen a shift towards expressing linguistic knowledge in ways which allow its verification and processing by formal means. Tools originating in mathematics, logic and computer science have been applied to human language to model its structure and functioning. Various aspects of different languages are being described within formally defined frameworks proposed by a number of interacting linguistic theories.

The proposals deal with various levels of linguistic description, starting from the level of sounds (phonetics) up to the level of meaning. Partial grammars and lexicons now exist for many languages within various formal frameworks and collections of linguistic analyses of text and speech are accumulated to be employed both in theoretical research and applications. Besides approaches relying on symbolic means and `rationalist' efforts which result in language models consisting of grammar rules and lexical entries, alterna

1 Grant of the Grant Agency of the Czech Republic No. 405/96/0198, which has now become an integral part of a newly awarded long-term grant of the same agency No. 405/96/K214
tive methods employ statistics computed from input text or its analysis to produce a stochastic model.2

However, a common and crucial issue cutting across all types of enterprise in this domain is the need to adopt or design an adequate formal representation of language structures in order to accommodate relevant linguistic knowledge in its relation to the actual language data. There is a number of tasks which typically require soundly defined formal representation of language structures:

1. analysis (parsing) of input text or speech into a representation, tagging of text or speech collections;

2. synthesis (generation) of output text or speech from a representation;

3. mapping of one representation onto another transfer (typically in machine translation systems).

These are the elementary tasks which are parts of many natural language processing applications, some of which are listed below:

machine translation systems;

natural language interface to knowledge bases, question answering systems;

automatic abstracting and knowledge acquisition systems;

automatic acquisition of linguistic data and its integration into a language model.

Formal representations of language structures which have been proposed by different linguistic theories and/or used in natural language processing applications reflect their context in many respects and suitable candidates for an intended more general use are difficult to find. This is due to various aspects of their design, such as (i) specific theoretical commitment, (ii) limited expressive power in partial coverage of language phenomena and restriction to certain levels of linguistic analysis, (iii) difficulties in expressing relationships between different levels of analysis, (iv) hard-wired reliance on some characteristics of a certain language or language group and the resulting difficulty in adapting the framework to a typologically different language,

2 When a linguistic description is implemented on computers, the usual goal is to parse sentences and produce representations of their analyses, thereby verifying the framework, the linguistic theory and the description itself. Another way to obtain (morphological and syntactic) analysis of sentences is by employing statistical methods on large samples of (already analyzed) texts in order to process a new text afterwards, performing some degree of linguistic analysis on the basis of the data acquired in the `learning' phase. Both these kinds of efforts converge and their increasing potential is reflected in the growing amount of text and speech data analyzed to a different degree for various purposes.

and, finally, (v) application-specificity. Thus, it is difficult to express a full-fledged syntactic analysis of a `free word-order' language by means of word-class labels and constituent brackets used for tagging (mostly English) texts.

Although it is not likely that a single framework could become a universally accepted vehicle of linguistic knowledge, we believe that a higher degree of generality and flexibility can be achieved for the benefit of both theoretical studies and application-oriented projects.

2 Characteristics of a Satisfactory Solution

From the conceptual point of view, an adequate design of formal representation should be able to express linguistic facts related to the following levels of description:

1. level of phonetics, phonology, graphemics: specification of phonemes, stress and prosodic patterns, etc.;

2. level of morphology: morphemes, morphological categories;

3. level of syntax: syntactic categories, syntactic structure (trees);

4. level of (linguistic) meaning: disambiguation of lexical meaning, specification of underlying structure and function, communicative dynamism and topic focus articulation, anaphora resolution.

There are several important features that should be reflected in the design to make it really useful:

It should be possible to describe a language structure in all its aspects simultaneously, i.e., to be able to relate facts from all levels of linguistic analysis in a straightforward fashion. At the same time, the design should permit access to specific aspects of the description without other aspects intervening. Thus, a user interested only in syntactic structure should be able to filter out any other information.

If a certain aspect of linguistic description can be structured and viewed differently depending on theoretical commitments, the design should provide an option to derive the desired way of presenting the linguistic facts from a common representation. Thus, both phrase-structure and dependency trees could be derived from the description.

The design should be capable of accommodating typologically different languages without substantial modifications, especially, it should provide space for stating the relation between word-order variations and higher levels and for the interplay between morphology and syntax in the case of complex expressions.

A related requirement concerns the possibility to express links between parallel structures and their analyses in different languages. This feature is important if parallel bi- or multilingual data are to be analyzed and studied as contrastive language structures.

The design should provide space for as little or as much linguistic facts concerning a language structure as is possible or practical to collect or express. This feature would permit to integrate text or speech samples with their analyses in a stepwise fashion, possibly starting with a bare text/speech string and leaving some levels unspecified.

It should be possible to represent at least some linguistic facts in an underspecified form. Wherever possible, an option to use a quantitative measure should accompany such cases. Disjunctions restricted to local domains, underspecified descriptions and weights could be the means to achieve this requirement.

The formal representation should be convertible to another format, as required by an application or desired by another specification covering compatible conceptual issues.

The design should be flexible in the sense that it should contain as few inherent restrictions to its extensions and modifications as possible.

3 Background, Methods and Problems

Without attempting to preview the results, the following points can be made to sketch the starting point situation, the outlines of the goal, and the path towards its achievement:

1. The project will be able to profit from theoretical results and practical experience gained in the field of formal description of natural language at our sites.

The fruitful results concerning word-order variations and their relation to meaning, as well as the richness of syntactic studies based on a dependency-oriented model, both widely acknowledged and faithful to the high standards of the Prague School linguistic tradition, provide a wealth of stimulating material.

At both sites, a number of application-oriented research projects have been at least in some respects tackling the problems of an adequate representation of language structures. The projects include machine translation, natural language interface to knowledge bases, automatic abstracting, automatic knowledge acquisition from texts and grammar checking.

2. The smallest piece of information (typically, a linguistic category) is expressed as an attribute and its value (i.e., a `feature'). A collection (conjunction) of such pairs is used to describe a linguistic object (typically an aspect of linguistic description of a word or a collocation), allowing for partial information (underspecification) and entering into more complex structures, where some attribute values are not atoms but structures. Through the recursive nature of such a representation, linguistic structures of arbitrary complexity can be described. Two or more attributes can share a single value, which is a possible way to implement relations between linguistic facts at different levels of description.

As structures of this type have become a kind of standard in modern linguistic research, the issues of compatibility with other approaches will be substantially simplified on many levels.

3. The design will be tested by its application on language data in at least two typologically different languages. A sample of bilingual parallel text data will be provided to test the parallel link option between analyses of linguistic structures.

There are a few challenging issues which call for an inventive solution:

The relation between the surface string of graphemes/phonemes, hierarchical syntactic structure and the ordering of meaning-bearing elements according to the degrees of communicative dynamism is far from straightforward. This concerns especially cases of crossing dependency (non-projective structures). If the representation is to accommodate descriptions on all levels in an integral form, a non-trivial solution has to be found.

C

Complex expressions like idioms, compound words and morphological categories realized by discontinuous sequences of auxiliary words present another problem of a similar kind.

The integration of all kinds of linguistic knowledge in a single formal framework capable of application to the widest range of language structures is a unique enterprise. Disregarding the undoubtedly immense practical profit for a moment, the project will probably bring the most precious theoretical fruit precisely in this domain.

4 The Treebank

The formalism developed within this project will be applied towards a mid-size treebank, mainly on the Czech material. There will be three layers in the treebank.

tation problem in such a complex and unified way. Also, the development of the past ten years will lead to novel approaches in the representation theory.

However, the idea of the ``development cycle" involving immediate, large-scale evaluation and verification on real texts has not been exploited previously in the framework of such a theoretical issue as a formal representation of language structures undoubtedly is. There are various projects, mainly in the United States, which do use the repetitive evaluation strategy to get valuable feedback, but they are more application-oriented. We feel that an appropriate modification and proper usage of such methods would mean a qualitative leap in a search for a theoretical result in a non-technical discipline. We would like to cooperate as much as possible with the centers doing a lot of work in this direction, namely, the LDC (Linguistic Data Consortium) at the University of Pennsylvania, and use their materials, especially for the evaluation phase of the English side.

There are also projects the results of which (or at least some of them) would help this project: this would also make very effective use of funds spent on other grants and research activites both within and outside of the Czech Republic. We envisage the use of some of the results obtained in the following projects: Grammar Checking for Slavic Languages (a PECO project, funded by the EU), from which we would like to obtain some ideas about representations of ill-formed input; Czech National Corpus project (funded by GAÈR), as a resource of Czech textual material; and MATRACE (also funded by GAÈR), as a starting point for comparison (and later, unification) of structural representations developed for the purpose of machine translation between two typologically different languages.

5 A Summary of the Goals

There are two main goals to be achieved:

A specification and thorough description of a single formal representation of language structure, integrating and enhancing the previous theoretical results, and adding new contributions at the same time (especially the representation of topic/focus, coreference, discontinuous elements relations, etc.);

An experimental verification of the above, i.e. the markup of a substantial portion of diverse, real text samples using the formal specification developed under the grant. In other words, building a treebank. Two typologically different languages will be used for the experiments, Czech and English.

We consider the two goals mutually indispensible, as we believe that only a rigorous testing of any formal representation theory will put it on a solid ground, and it will make an immediate feedback possible.

References

[1] Petr Sgall, Alla Goralèíková, Ladislav Nebeský and Eva Hajièová, *Functional Approach to Syntax*, American Elsevier, New York, 1969

[2] Petr Sgall, Eva Hajièová and Jarmila Panevová, *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*, D. Reidel Publishing Company, Dodrecht, 1986

[3] Vladimír Petkevic, *A New Formal Specification of Underlying Representations*, Theoretical Linguistics 21, 7-61, 1995

Syntactic Tagging at INL

Stephan Raaijmakers

Institute for Dutch Lexicology (INL)
Leiden, The Netherlands

INL annotates large text corpora with PoS and lemma information, using rule-based and stochastic taggers/lemmatisers. For the application of PoS-tagging, syntactic analysis can be quite useful: it may establish locality between an ambiguous PoS and its resolvent, allowing locally operating models (such as Hidden Markov Models) to resolve the ambiguity. At INL, some exploratory investigations into syntactic tagging are being carried out, at this moment primarily for the purpose of improving on PoS-tagging. The investigations address the problem of grouping the context between PoS-ambiguity and resolvent into constituents.

Two `classical' parsers have been implemented: a CYK (chart) parser, and a deterministic shift-reduce (Marcus) parser. No large grammars have been written for these parsers, yet. The parsers are being used to study the intertwining of syntactic knowledge with the PoS-disambiguation rules of INL's rule-based tagger/lemmatiser DutchTale.

An alternative approach, boundary marking, produces shallow syntactic representations without fine-grained internal structure: it generates top-level phrasal boundaries, like in:

- [The student]-[will buy]-[the cheap edition].

Boundary markers do not need large grammars. A prototypical boundary marker has been implemented, using a small set of boundary-placing rules. It is unclear yet whether PoS-tagging needs to address syntactic structure of a higher sophistication than the shallow structure produced by boundary markers.

Contrasting with these approaches, self-organising models are investigated as well. A backpropagation neural network, at the moment being used at INL for morphosyntactic disambiguation, can be trained on context-free grammar rules, and can be supplemented with tree construction routines to behave like a parser. It is possible to train the net on a relatively small core grammar, and let the net produce creative solutions for patterns outside the coverage of the training grammar (robustness).

A radical solution to the problem of writing large grammars for syntactic tagging will be the use of self-organising maps (SOM's), which can be used to construct a topology of syntactic clusters without prior formulated linguistic knowledge. These clusters can be interpreted as syntactic categories. This will be a topic of interest in the near future.

Editorial | Syntactic Tagging | Awareness Day in Bucharest | Birmingham Workshop

News from TELRI Working Groups

- ***WG2 DOCUMENTATION***

  Co-ordinator: Ruta Marcinkeviciene

  Since the October meeting 1996 of the participants from three projects: the ELSNET, ELSNET goes East and TELRI, the two latter projects have joined their efforts for documenting Eastern European NLP and Speech sites for the sake of a greater efficiency and cost reduction in carrying out their mutual tasks.

  During the firts half of 1996 a revised joint questionnaire meeting the needs of all three projects was prepared and sent out to both Western (by ELSNET) and Central and Eastern European (by ELSNET goes EAST) countries. The questionnaires were sent out both by e-mail and surface mail from Amsterdam with the hope to have a slightly increased rate of response. 249 questionnaires were sent out to 11 countries: Baltic countries, Belorus, Bulgaria, Czech Republic, Hungary, Poland, Romania, Slovakia and Slovenia. 167 were distributed by e-mail, the remaining ones by surface mail. By the end of March about 50 of them came back answered to Amsterdam and keep coming all the time. Most answers come from the e-mail sites. The greatest percentage of answers came from the Baltic countries, Czech Republic and Poland.

  TELRI WG 2 actively participated in the creation of the new joint questionnaire with the aim of careful documenting of language and speech resources according to the accepted pattern. We supplied the list of addresses of language and speech engineering organizations with 43 addresses mostly from those country which participate only in one of the projects, i.e.TELRI. Now TELRI participants are responsible for those completed questionnaires which come by the surface mail. The next task for both projects is to prepare European NLP and Speech Survey in electronic and paper versions.

- ***WG10 USER NEEDS***

  Co-ordinator: Andrej Spektors

  The aim of WG 10 for further period will be working out proposals of projects dealing with computational methodology and software for semi-automatic validation of the corpora and lexicons. At present there are no strict standards adhered to by all resource developers, although no one objects a standard adoption. Existing standards and guidelines developed while working on various projects are mostly used on the level of recommendations, and resource developers do not always observe them to full extent. It has to be noted that development of linguistic resources in Central and Eastern European (CEE) countries still is in its initial stage, therefore timely introduction of already developed standards during the course of resource creation would be beneficial, resulting in a considerable economy of financial resources. Of course, any standard can be introduced only by gradual acceptance by Natural Language Processing community. Therefore already accepted and validated standards have to be offered.

The lack of appropriate tools for validation of written language resources constitutes a serious impediment to wide-scale commercial exploitation of these resources. Prospects of introducing semi-automatic methods for resource validation in practice are especially good in CEE countries where creating of resources in national languages is in the initial stage. The previous experience shows that existing resources there mostly do not conform to standards or specifications and are not harmonised in content and form. Existing resources in CEE countries most often are set up for internal use of producer's institution and are not commercialized. Semi-automatic validation of formal properties would facilitate distribution of all written language resources built up in accordance with SGML and TEI formalisms.

Exploitation of linguistic resources in CEE countries at present is in an early stage, therefore timely development of methodology for validation of created resources is of utmost importance, providing grounds for minimizing financial resources necessary for error elimination and standardisation in the future. Practical usability of standards and recommendations for different languages, which are of more inflected nature than English and other Western European languages, would be tested during realisation of such further projects.

Tools for verification in accordance to standards will be designed to create the necessary means for testing a correspondence of language resources to as many existing standards and recommendations as possible. The possibility to add tools for testing resource compatibility to new standards and recommendations in the future will be supported. Development of tools will start with collection of information on all participant national languages and with a co-ordinated evaluation of this information. After information collection and evaluation the experimental software will be created and tested for all national languages. Possibilities to reduce other tagging methods to the SGML standards will be inspected.

National language engineering standard centers will be established in WG 10 participants' countries where interested persons of the country will be able to study existing standards, specifications, recommendations and evaluation methods in computer linguistics. During the work on theabove-mentioned projects the participants have to study all standards in detail. Therefore standards (guidelines, recommendations, specifications) have to be collected together, and the requirements have to be carefully studied. Specific recommendations for use of standards and specifications for the corresponding language will be developed in these standard centres.

A proposal for checking of correctness according to SGML and other recommendations will be created. First statistics about which and how many SGML tags are used and all other possible statistics about resource tagging will be collected. Possible usage of these statistics for automated resource evaluation will be investigated. Such statistics will be collected by each WG 10 participant for the respective national language and algorithms will be developed. Methods and algorithms will also be developed to test correspondence of language resources to TEI and EAGLES recommendations and specifications developed by PAROLE project.

Syntactic Tagging | News form TELRI Working Groups | Awareness Day in Bucharest | Birmingham Workshop

Awareness Days in Bucharest

Dan Tufis

58

1. General comments on the Seminar organisation

During January 29-30 1996, the Romanian Academy (Center for Advanced Research in Machine Learning, Natural Language Processing and Conceptual Modelling) organised in Bucharest, the National Seminar "Language and Technology", fully funded by the European Commission under the programme "Awareness Campaign on Language Technology". The organisation of the Seminar has been supported by the Department for European Integration of the Romanian Government and the Research Institute for Informatics.

The National Seminar was a very successful event, being attended by more than 250 participants, from research, industry and public administration. Policy making sector was represented by high level representatives. Public administration was also represented by several head of departments in key ministries. Some big private companies in Romania were represented by their directors. Big state industrial and development companies had a significant representation in the audience and in the scientific program. Different SMEs expressed their interest in the Seminar not only by taking an active part in the event but establishing contacts aiming at finding possibilities for marketing some of the systems that were demonstrated during the Seminar.

Academic community represented more than 50% of the Seminar participants and most of the Romanian representative scholars attended the Seminar. Most of them came from the field of traditional linguistics and philology, but computer scientists, mathematicians and cognitive scientists were very well represented, too.

The Seminar had a real national character, being attended by people representing all important university towns of Romania: Bucharest, Iasi, Cluj, Timisoara, Craiova, Constanta, Sibiu, Brasov. The Seminar was largely advertised in mass-media. There were press announcement, published in the nation-wide newspapers and weekly journals (Academica, Economistul). The Chairman of the Organising Committee gave three pre-seminar interviews on the national radio broadcasting programs. During the Seminar days, more than 15 persons (including EC officials) were interviewed. Two popular TV broadcasting companies included images and comments on the Seminar into their news.

2. The Seminar Program

The Seminar lectures were given in the Magna Aula of the Romanian Academy, the most prestigious conference room in Romania. The demonstrations were given in the Presidium Room, next to Magna Aula, specially equipped for this event with a heterogeneous local network (5 Pentiums, 4 IBM486, 1 SUN Sparc 4, and 2 Macintoshes). For the entire period of the Seminar, ear-phone simultaneous translation between Romanian and English were ensured by specialised translators. The work they done was gratefully acknowledged by both the organisers and the participants of the Seminar. Due to the initial intention to have the contributions to the Seminar published by the Romanian Academy Publishing House, the most prestigious publishing house in Romania, a reviewing committee was formed and all the submitted contributions (except for the invited talks) were independently reviewed. Out of the 43 submissions 29 papers were accepted for presentation. The volume, which included also the 12 invited papers, is considered to be quite representative for the state of art in Romania as far as language technology addressing Romanian is concerned. The Seminar was opened by the Vice-President of the Romanian Academy, Professor Aureliu S`ndulescu. Professor Marius Guran, presidential advisor on science and technology, presented the salute message on behalf of Ion Iliescu, President of Romania. Secretary of State Dr. Ghiorghi Pris`caru, Head of European Integration Department of the Romanian Government, presented a salute message on behalf of the Romanian Government. Mrs. Karen Fogg, Head of the European Commission Delegation in Bucharest presented a warm salute from the European Commission, highly appreciated by the audience. Secretary of State, Mircea Petrescu, President of the National Commission for Informatics, gave a keynote speech on the informatising strategy in Romania. The second keynote speech was given by Jan Roukens from the European Commission-DGXIII, on one of the hottest issues of our present-day society: "Breaking the Language Barrier: Towards a Multilingual Information Society in Europe".

After the Opening Session, there were 3 sequential communication sessions:

Language Resources, Machine Translation and Speech Technology and in parallel there were several demonstrations with language technology systems implemented in Romania and addressing Romanian language.

Five invited talks were given on the first day of the Seminar:

Wolfgang Teubert (IDS-Mannheim) _ "Language Resources for Language Technology"

Svetlana Cojocaru (Academy of Sciences of the Republic of Moldova) _ "Romanian Lexicon: Instrument, Implementation , Use"

Walther von Hahn (University of Hamburg) _ "Machine Translation "

Rajmund Piotrowski (University of Sankt Petersburg) _ "Machine Translation in New Russia"

Peter Roach (University of Reading) _ "Speech Technology"

The first day of the Seminar was concluded with a round table on "Bridging the Gap between Theoretical Linguistics and Linguistic Engineering" (moderators E. Simion and M. Guran) with panelists from both communities: Wolfgang Teubert, Rajmund Piotrowski, Marius Sala, Alexandra Cornilescu and Marian Papahagi, Peter Roach, Walther von Hahn Alfred LeJia and Dan Tufis. For one hour and a half the panelists tried to analyze the existing gap between the researchers of the two disciplines and pleaded for a synergetic action in the benefit of language technology. The role of the education was emphasised as a key factor in bridging the gap and there were reports on some progress in this respect. The Technical University of Bucharest, Cluj and Iasi (the Computer Science Departments) included into their curricula courses on natural language processing and linguistic theories (HPSG, GB). The philological faculties (University of Bucharest, University "Babes-Bolyai" in Cluj) included in their curricula optional courses on text processing and computational linguistics. The program of the second day contained two sequential sections:"Applications: Research, Industry, Users" and "International Cooperation", followed by a round table on the topic "How could the international cooperation help the technology of Romanian language" with EC representatives and Romanian decision makers as panelists.

There were 5 invited talks in the two sections:

Gabor Proszeky (Morphologic, Budapest) _ "How to Reach the LT Market ?"

Poul Andersen (EC-DGXIII, Brussels) _ "Cooperation with Central and Eastern Europe; The European Commission's Strategy""

Steven Krauwer (OTS, Utrecht) _ "European Cooperation: The ELSNET Experience"

Eva Hajièova (Charles University, Prague) _ "Natural Language Processing in Czech Republic: National

Projects and International Cooperation"

Tomaz Erjavec (Josef Stefan Institute, Ljubliana) _ "International Cooperation in Slovenia"

The technical program of the second day of the Seminar was concluded by the round table "How could the international cooperation help the technology of Romanian language" (chaired by D. Cristea). The panelists were Jan Roukens and Poul Andersen from the European Commission and Marius Guran, Mircea Petrescu, Florin Teodor Tonisescu and Eugen Simion from key governmental institutions of Romania.

After some comments by the Romanian decision makers on the necessity for further concerted actions from the local institutions towards a more focused R&D activity in the field of language technology and statements concerning governmental support, the EC officials resumed some key principles concerning the international cooperation emphasising the need for openness and distribution of tasks. Several questions were raised from the audience, which were answered by the panelists. The Seminar ended with some concluding remarks made by Marius Guran, Mircea Petrescu and Jan Roukens. All the three speakers appreciated the National Seminar "Language and Technology" as a very significant event for the Romanian scientific community which was well managed and expressed their hopes for positive and synergetic follow-ups of the event. Special thanks were addressed by the Romanian Officials to the European Commission for making possible the Awareness Seminar in Bucharest.

Acknowledgments

Besides the European Commission, different individuals must me mentioned as recipients of our gratitude.

During the preparation of the Seminar, the organisers benefited from the assistance of Mrs. Grazyna Woszcieszko and Mrs. Helene du Callatay. Their readiness, fast and precise answers to the issues raised during the organisation of the Seminar were extremely supportive. Special thanks are due to Mr. Poul Andersen who deeply involved himself in preparing the Seminar (it suffices to mention a dozen of calls, more than 100 e-mail exchanges and three face-to-face thorough discussions on different meeting occasions). Besides his extremely useful experience, his patience and understanding are warmly acknowledged. The invited speakers delivered high level presentations, carefully prepared. Their efforts are sincerely acknowledged here.

Editorial | Syntactic Tagging | News form TELRI Working Groups | Awareness Day in Bucharest

TELRI Birmingham Workshop Report

Primoz Jakopin, SLOVENIA

INTRODUCTORY NOTES

The first TELRI workshop took place at the University of Birmingham in the week from October 10 to October 13, 1995. The TELRI Steering Committee accepted my application for a short term visit and so I could attend the event, which can rightly be described as most useful.

The workshop took place at the Corpus linguistics department of the School of English and at the COBUILD institution. There were 8 participants: Barbora Hladka from Prague, Ruta Marcinkeviciene and Vytautas Zinkevicius from Kaunas, Madis Saluveer and Tiit Roosmaa from Tartu, Ana Maria Barbu and Maria Lidia Diaconu from Bucharest, and myself. We all have known that the Birmingham corpus of

English texts, Bank of English with 210.5 million words, is the biggest existing, but to see it and its use on the spot is very different from knowing it only from literature.

We were also very pleased by the warm reception and all over hospitality of our host, Prof. Dr. John M. Sinclair (JMS) and of his team. They spared no effort to help us with our task, to see the essential working and benefits of such corpus in the span of a few days. Lectures were accompanied by rich descriptions on paper, including examples and Prof. Sinclair generously provided everyone of us with several books, including the New COBUILD Dictionary of English. As most of us arrived in Birmingham a day earlier, the University library proved its worth on Monday. It is well stocked in the field of computational linguistics and I could also find a lot of new foreign titles, most notably German, such as the ones from the QUANTITATIVE LINGUISTICS series.

The stay at the Lucas house, only a short walking distance from the Department of corpus linguistics, was also very agreeable. The institution of English breakfast was new to me and it surprised with its variety and richness; especially as I came with false preconceptions that the English food is mainly limited to fish and chips. Even good weather, for the lack of which the Island is well known, contributed to the success of the workshop. It kept throughout and from a rented bicycle I could even catch a glimpse of the Birmingham countryside, with its vast network of navigable water canals from the end of 18th century, lately furnished with sidewalks. It was interesting to see how old can be put to good use at the present time, and the suggestion of Prof. Sinclair that the quickest way to get from the University to the very centre of Birmingham is by the canal sidewalk, proved very accurate. It was 11 minutes by bike.

THE WORKSHOP

The workshop started on Monday with a reception in the Westmere main building. After some introductory words by Dr. Wolfgang Teubert, head of the TELRI project, and by Prof. Sinclair, there was an opportunity to discuss matters with workshop teachers and the people from Cobuild. The remark of Ramesh Krishnamurthy, corpus manager at Cobuild, that the lemmatisation of languages with rich inflection, such as Slovenian, should be easier than that of English, as there is more information for the mechanism to catch on, was highly interesting and provocative.

On Tuesday morning Prof. Sinclair gave an overview of corpus linguistics, from the first beginnings at the end of the sixties to recent achievements, such as the Bank of English, more efficient teaching, new ways of looking at the phenomena of language and better dictionaries, all coming out of it, to what can be expected in the future. Elena Tognini-Bonelli followed with an interesting, fresh approach on how corpus data, especially the collocations, can be put to good use in resolving ambiguity problems and proper use of words in translation. It was illustrated by examples in English-Italian context; as the summer school of Czech language in Prague also taught me quite some Italian, I enjoyed it very much. In the afternoon Tim Johns, who is involved in teaching English for the International students unit (2.000 students) at the University of Birmingham (12.000 students in all), described the concept of data-driven learning. The accompanying teaching material on paper and his own software, CONTEXTS, showed how to teach a language in an enjoyable, yet very efficient way. From the lecture it was very clear that the work of JMS had taken deep roots in Birmingham; corpus-driven learning is no novelty there.

COBUILD

Wednesday morning was devoted to the visit of Cobuild, COllins Birmingham University International Language Database (acronym invented by JMS), a joint venture between academic (Univ. of Birmingham) and industrial (Harper-Collins) partners. The University expected support in building a large-scale text corpus while the other side expected increased competitive edge through better dictionaries, with entries and explanations selected by their actual frequency and not at the liberty of dictionary authors. The project started in 1980 with a 50:50 share and pushed on with substantial funding from Collins: 1.5 mil. GBP from 1980 to 1984 and additional 1 mil. GBP from 1984 to 1987 made Cobuild the largest joint project in humanities worldwide. The data base grew from 7.3 million words in 1983 to 211 mil. now in the Bank of English and the staff to 20 full-time employees of today.

On the hardware side the work started on the University's ICL 1900 mainframe in 1980 and expanded in 1982 with the purchase of a DEC PDP 11/34 minicomputer (256 KB of RAM, 134 MB on disk), their first machine with UNIX (MULTIX) operating system. Independence from the University mainframe was achieved in 1987 through the own network of RISC workstations (IBM 6150). A network of PCs was considered but dropped due to the belief that PCs would not be up to the task, while the workstations, though much more expensive, would. It would be interesting to see what the decision would be today, as the margin in capability between high-level PCs and workstations is vanishing fast. The Cobuild network has been upgraded to Sun-Sparcstations (2 servers and 18 diskless workstations) and Tektronix terminals (16, with 17 inch screens) in 1991. The software used at the beginning was the concordance builder COCOA by Atlas, supplemented by own software (XLOOKUP) after 1983. Later in the course of the project Collins publishing house was acquired by Rupert Murdoch, the media entrepreneur, who also wanted greater control over Cobuild. His Harper-Collins now owns 75% of Cobuild; the 25% share of the University however excludes it from vital decisions. The good side of the new parent arrangement is that Cobuild can get access to all the publications from the media empire just mentioned, such as the newspapers Independent, the Times, Economist, New Scientist and Guardian for free. Bank of English (BOE) contains 75% of material in British English and 25% in American. It does not include poetry, drama or child language.

The program XLOOKUP, which is used to retrieve data from BOE, indeed performs impressively. It can also be tested, with limited access to a subset of BOE, 20 mil. words, via Internet. The relevant electronic address, ID and password are: *titania.cobuild.collins.co.uk login: cobdemo password: cobdemo*

Such a tool at one's disposal augments the possibilities of most research in the field of English language by an order of magnitude. The online access to collocational information on words is also of immense value for anyone writing in English.

OTHER SOFTWARE

Wednesday afternoon was devoted to concordance and collocational software (WordSmith Tools) on standard desktop computers, PCs, written and presented by Mike Scott from the University of Liverpool. The software, intended for lexical analysis on PCs and for studying the output of larger computers on smaller machines, is planned for publication by the Oxford University Press. It runs on Windows environment, is quite impressive and reflects the author's great experience in the field.

On Thursday the home-grown software tools, from the Department of corpus linguistics, were shown and demonstrated by Oliver Jacobs. Due to the large pressure put on the Cobuild staff in the new circumstances, to make as much marketable output as possible in the form of new dictionaries, the needs of the academic side (with its smaller share in the company), evidently had to be put aside in the development of XLOOKUP. The necessity for the Department to have its own software has become urgent and will be met in the next several months. It was however interesting to see how, in the world of workstations, PCs are inevitable as well. All the demonstrations were performed on PCs serving as UNIX-terminals of larger workstations; the reason seemed to be the lack or unavailability of LCD overhead projecting facility on workstations.

CONCLUSION

Triple C word, title of the famous book by JMS: Corpus, Concordance, Collocation, had, for most of the participants of the workshop, only terminological value. In Birmingham we all got an understanding of what it takes to construct a real textual corpus, to maintain it and how to exploit it fully once ready.

To compose the Bank of English was no straightforward task and many temptations had to be avoided. One of the most difficult points in such considerations, especially if the size of the data base is expected to grow from megabytes to gigabytes, is what to do with errors - typos and the like. Prof. Sinclair's answer to the question was highly illustrative: "Errors are part of text. If you correct them, you lose information." His other remark, on what words to include in the future dictionary and which not, is also worth noting here: "If a word has a frequency of more than 15 in your corpus, you must have very good reasons not to put it in; if less than 15, for including it."

There are three other points worth of further consideration:

1. The XLOOKUP program would benefit greatly, as I see it, from a graphical user interface (GUI) it now lacks. The proportional screen fonts would allow much wider word surroundings and the colours would help the collocations, especially non-adjacent ones, to stand out better.

2. In addition to displaying the current state of English, the Bank of English increasingly has an encyclopaedic value. It could prove very useful and would attract much wider academic and non-academic audience, if the Bank included all collected material and not only the current one. It would be technically feasible, even now already, to have a data base larger by an order of magnitude, at least ten times. It should be accessible to inland users and the world community via Internet and be housed in an institution with similar status and funding now characteristic for the National library.

3. I also missed very much a good statistical description of the Bank of English, on the character, word and sentence level. In the short time of the workshop it was not possible to obtain the word frequencies I would need for the plotting of rank-frequency curve which I would like to compare with data from smaller samples. It is my great hope that this would be possible in the not-so-distant future.

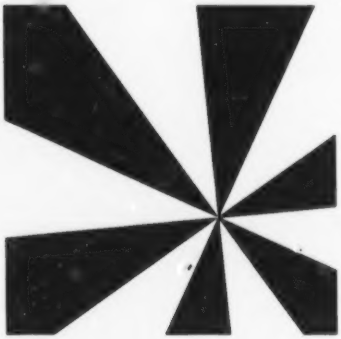
All in all, the knowledge gathered in Birmingham widened my horizons very much. Together with the overview which I obtained during a visit to Institut fuer deutsche Sprache in Mannheim two years ago it will help a great deal in any similar future task for the Slovenian.

Top of this issue TELRI Main Page IDS Main Page

TELRI

TRANS EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE

Concerted Action in the Framework of the Copernicus Program



Newsletter 4 *October 1996*

Contents:

1. Editorial ---------------------------- 3
2. Topic of this issue: Syntactic Tagging (continued) - - 6
3. The Czech National Corpus --------------- 28
4. TELRI Event - Nancy TEI Workshop --------- 30
5. New prospective member of the TELRI advisory board 32
6. Some interesting events - past and future ----- 35
7. List of Participants ------------------- 38

Coordinator:

Dr. Wolfgang Teubert
Institut für deutsche Sprache
P. O. Box 101621
D - 68016 Mannheim, Germany
phone: +49 621 1581 437
fax: +49 621 1581 415
e-mail: telri@ids-mannheim.de

Editors:

Prof. Eva Hajičová
Mgr. Barbora Hladká
Institute of Applied and Formal Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25
118 00 Prague 1, Czech Republic
tel.: +42 2 21 91 42 88
fax: +42 2 21 91 43 09
e-mail:
hajicova@ufal.mff.cuni.cz
hladka@ufal.mff.cuni.cz

Contributions to TELRI newsletter, and address corrections, should be sent to:

e-mail:
hladka@ufal.mff.cuni.cz
fax: +42 2 21 91 43 09

Editorial

Wolfgang Teubert, *Coordinator of TELRI*
Norbert Volz, *Project Manager*

1. OUR PRESENT POSITION

The TELRI Plenary Meeting held in Mannheim, June 14-18, 1996, marked not only the middle of the project's timetable but also the turning point from TELRI as a network under construction to a functional and highly successful pan-European language resources infrastructure. Our external reviewers gave us a favourable evaluation of the performance and motivation displayed by all partners – an achievement we can duly be proud of, but also an obligation to maintain this positive image for the future.

The encouragement received from our external evaluators has shown us that it is now time to forge the link between research and infrastructure activities as agreed upon at the previous Steering Committee Meeting.

2. CHANGE OF WG STRUCTURES

Facing these new challenges, we have decided to restructure the scope, membership, and coordination of TELRI working groups, especially in the network and service area, and to change the ratio between infrastructure-oriented and research-oriented Working Groups.

As a first step towards this aim, WG "Seminars" was changed to WG "Morphosyntactic Annotation", and WG "User Needs" was transformed into WG "Validation". Also, a survey was held among all TELRI partners in order to collect and identify further needs for changes in Working Group scopes and structures and to subsequently set up a new Working Plan for the next period of the project.

The Joint Meeting of the WGs Joint Research, Lingware Availability, and Networking at the Mannheim Plenary Meeting and the Nancy Workshop on Service Tools, August 28 - September 1, have also shown that there is a strong demand for closer cooperation and joint activities between the members of these Working Groups that could be beneficia to the project as a whole. Further joint activities will include a Workshop on Public Domain Software Tools as well as continuing and expanding the work on the Plato Parallel Corpora and Birmingham's COBUILD Bridge Dictionary project.

3. ACCOMODATION OF WORKPLAN ITEMS

For the end of 1996, TELRI plans to prepare a proposal for a new COPERNICUS project "Multilingual Terminological Database" along with international partners such as ISO, Infoterm, or ELRA. This will serve as a repository for the terminology of Language Engineering and Language Technology worldwide.

Further activities in the area of documentation of LR/LT activities and resources include a bibliography on corpus linguistics, to be completed by the IDS early next year, and the establishment of a searchable database for TELRI resources on the WWW, where a prototype version has already been installed and can be accessed via the TELRI Website.

Working Group "Joint Research" (Coordinator: John Sinclair, Birmingham) is continuing its work on the localisation of the COBUILD Student Dictionary to the various TELRI languages. In an electronic form, these dictionaries can easily be linked and, thus, serve as a multilingual lexicon.

The newly constituted Working Group "Validation" (Coordinator: Primoz Jakopin, Ljubljana) will, in cooperation with the ELRA Written Language Resources Validation Panel, establish an analytical framework for the validation of non-SGML corpora and tools.

TELRI will prepare a proposal for a workshop on Morphosyntactic Annotation, to be held in Spring 1997. Participants will have the opportunity to discuss annotation standards and recommendations with experts from EAGLES, MULTEXT, PAROLE, and other related projects in order to adapt existing guidelines for morphosyntactic annotation to suit the peculiarities of the non-EU-languages represented in TELRI as a further step towards the development of comparable pan-European language resources.

For June 1997, TELRI plans a workshop on Translation Equivalents, to be organised jointly by WG Joint Research and the Tuscan Word Centre. For this activity, links will also be established with the PAROLE Association Working Group "Multilingual Lexicon".

Also for 1997, TELRI is going to launch the proposal for a new project "PAROLE East" to be submitted under the next COPERNICUS call for proposals. Complementary to the existing LE-PAROLE project, PAROLE East aims to create standardised language resources (comparable corpora and lexicons) in those CEE and NIS countries that already have some resources available for conversion according to existing PAROLE standards.

4. THE FUTURE OF TELRI

In their evaluation report, our external reviewers, Prof. Alexander Barulin (Moscow) and Dr. Mark Liberman (Philadelphia) strongly supported TELRI's intention to set up an independent legal body, "TELRI Association", in order to enable the continuation of our activities once the present funding period has expired. We are now in the process of establishing the TELRI Association as a registered association under German civil law, which we hope will be finished by the end of this year.

In response to another desideratum expressed in the Evaluation Report, TELRI will continue to expand and improve its presence on the World Wide Web. As a first step, we have established a dedicated position of TELRI WWW Officer and have contracted this job to a graduate student who will maintain and update our WWW pages. In addition to the already existing English WWW pages, it is planned to set up individual pages for each of the TELRI languages as well as to install "interactive" WWW pages for quick and easy information on TELRI resources and services. We thus hope to make TELRI not only a viable and successful institution but also a long-term "brand name" on the corpus linguistics information market.

Topic of this issue: Syntactic Tagging

(continued)

***NOTE OF THE EDITORS:** We bring two more contributions to the discussion on tagging started in the preceding issue of TELRI Newsletter. Since we believe that corpus annotation belongs to the hot problems in the present state of development of corpus linguistics, we would welcome any reports on ongoing projects, proposals of innovative approaches or comments on the already published contributions.*

Morphosyntactic Annotation of Textual Corpora using the LE-PAROLE Tagset Specifications

Norbert Volz, IDS Mannheim

1. INTRODUCTION

According to Leech and Wilson (1994, p.3), "*Corpus annotation* is the practice of adding interpretative, especially linguistic, information to a text corpus, by coding added to the electronic interpretation of the text itself."

Thus, *morphosyntactic annotation*, also known as *part-of-speech (POS) tagging*, is the annotation of the grammatical class of each text token. POS tagging is primarily carried out automatically by using rule-based or stochastic (e.g., Hidden Markov Models) tagging algorithms (see e. g. Hladká and Hajič, 1995). Therefore, we will concentrate on machine tagging of corpus texts, either fully automatic or semi-automatic, i.e., involving human intervention at some stages.

Within large multilingual projects such as PAROLE or TELRI, most of the available corpus tools such as taggers, access, maintenance and storage software, etc. are developed at different locations and are often also based on different existing resources and programs; therefore, common encoding standards and guidelines have to be developed in order to port the various tools and resources to other partner sites in the project.

2. "TASK-ORIENTED" AND "RESOURCES-ORIENTED" APPROACH

When it comes to the actual design of a tagset, two basic approaches can be distinguished with regard to the scope of the envisaged application. To make this distinction more clear, I will describe these two different *modi operandi* as either "task-oriented" or "resources-oriented".

2.1 "TASK-ORIENTED" OR "ECONOMICAL" APPROACH

A *task-oriented approach* aims to produce the maximum level of morphosyntactic annotation in the most economical way with those resources and tools that are readily available at present. It is mainly used if the tagged texts will serve as input to some concrete application such as context analysis or translation software, where the amount of morphosyntactic information available is enough to fulfill the requirements of the superordinate task. In other words, morphosyntactic features that cannot be identified by the tagger will not be included into the tagset; also, the number of ambiguities left to be resolved will be reduced as the tagset inventory is more or less restricted to those features known to be available and unambiguous. The advantage of those "stripped" or "poor" tagsets ("poor" meaning only having a small number of tags Ł 100) lies in their rather high tagging success rates of around 96% (Erjavec 1996). This advantage, however, is paid for with a certain indistinctness and lack of flexibility of the tagset.

2.2 "RESOURCES-ORIENTED" APPROACH

A *resources-oriented approach* aims mainly at the creation of large generic corpora. These annotated corpora not only serve as state-of-the-art material for today's applications, but, in the form of "reference corpora", serve also as a textual basis for future research. Therefore, it is desirable to reach a maximum level of morphosyntactic annotation, i.e., as detailed and fine-grained as the lexical encoding. Of course, this "ideal" level cannot yet be achieved with the automatic tagging algorithms available at present and, therefore, still requires a fair amount of manual intervention (tagging, checking, disambiguation) if carried out to its full extent. The aim is not only to accomodate the annotation level feasible at present but also to allow for further refinements and the inclusion of additional features. This requirements call for an "open" or "fringed" tagset structure, resulting in large, "rich" tagsets with sometimes more than 1800 possible combinations of obligatory tags (Ridings 1996).

3. MULTILINGUAL TAGSET DESIGN: THE PAROLE APPROACH

3.1 "COMMON CORE" TAGSET

The notion of a "common core" tagset is based on the EAGLES three-level distinction of *obligatory, recommended*, and *optional* features. The obligatory "minimal tagset" encompasses all morphosyntactic features considered as common to all involved languages. The "common tagset" includes additional features pertinent to most languages, and whose annotation is recommended (but not mandatory) for various reasons. The third, "optional" level is realised by language-specific extensions in order to cover the singular features of each language. Common core or "skeleton" tagsets are mainly applied by projects like MECOLB that use a task-oriented approach according to the definition given above. (Cloeren 1995, p. 3)

Figure 1, taken from Volz and Lenz (1996, p.1) shows the typical design of a common core tagset.

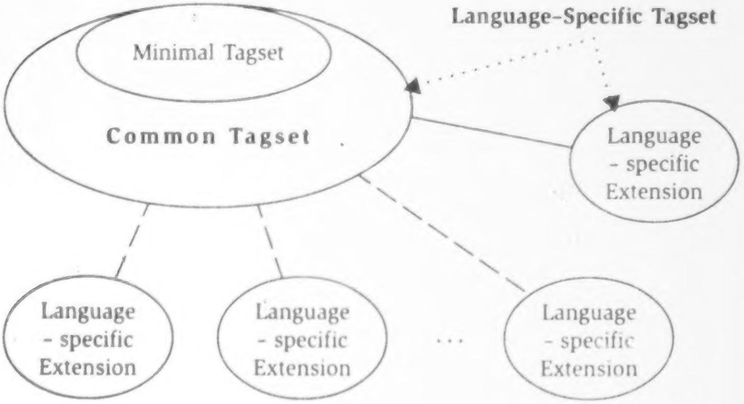

Language-Specific Tagset
Minimal Tagset
Common Tagset
Language - specific Extension
Language - specific Extension
Language - specific Extension
. . .
Language - specific Extension

Fig. 1. "Common Core" tagset design

However, the fixed character of a common tagset following the above principle will imply large language-specific extensions to make full use of the available tagging algorithms, especially for the languages outside of the EC language group for which these EAGLES recommendations were designed in the first place.

3.2 THE PAROLE APPROACH

The PAROLE tagset specifications are also based on the EAGLES recommendations formulated by both the Corpus and Lexicon Working Groups. The main idea was to provide a resources-oriented, multilingual tagset allowing for a high granularity of annotation.

It became clear, however, that designing a generic and portable, yet very fine-grained tagset across a variety of languages is not a feasible task, because both language-dependent and theory-dependent decisions have to be taken when mapping lexicon mark-up inventories to common corpus tagsets.

The main focus of specification standards within PAROLE has, therefore, been on the *notation* and overall *tagset design* rather than on linguistic criteria for the annotation itself. The idea is that a well documented and feasibly designed tagset will enable an adequate exploitation of the corpus under "real-world" conditions, that is, with imperfect data and with some theoretical aspects still unsolved, but able to accomodate future refinements and extensions. The resulting tagset specifications are thus based on the present state of the art, but not restricted to it.

PAROLE has therefore applied a "mosaic" approach where the distinction between "generic" and "language-specific" features is apparent within the tag proper rather than being an inherent feature of a particular "sub-class" of the tagset.

Basically, three steps were necessary to design the common tagset specifications:

- Collection and comparison of the categories and features required by all PAROLE languages.
- Establishing a general notation convention that allows for a gradual distinction between common and language-specific features.
- Definition of the hierarchy between features.

3.2.1 COLLECTION AND COMPARISON OF THE CATEGORIES

The selection of categories and features follows the EAGLES-based PAROLE specifications for the lexicon as described in the Appendix 1 of the LE-PAROLE contract: *Lexicon Architecture and Model* and the specifications given in the MLAP-PAROLE report "Task 4.2.2: Lexicon: Morphosyntactic Specifications: Language Specific Instantiations", Pisa 1996.

3.2.2 GENERAL NOTATION CONVENTION

A predefined number of attribute positions 1 to *x* is kept for all languages. The number of positions refers to the common features shared by at least two languages in PAROLE. Language-specific or optional features that do not correspond to these positions can be included in the tag by using positions *x+1* to *n*.

Example: General notation convention for nouns:

	Common					**Optional/ Language-specific**
Position	1	2	3	4	5	6 ... n
Features	PoS	Type	Gender	Number	Case	e.g., Contrast
Attributes	Noun	common proper	masc. fem. neuter	singular plural	Nom. Gen. Dat. Acc.	marked unmarked

The German noun "*(des) Hundes*" ("the dog's/of the dog") would be tagged as follows:

PoS:	Noun	→ N	→ **Ncmsg**
Type:	common	→ c	
Gender:	masculine	→ m	
Number:	singular	→ s	
Case:	Genitive	→ g	

The equality symbol "=" is used for an attribute that is not tagged within a certain language tagset although present within the lexicon. The hyphen symbol "-" is used for features that are not applicable for a specific combination of attributes and values, e.g., if the attribute does not apply to a particular category subclass whilst still applying to the category as such. The hyphen is also used for features not applicable to a particular lexical item although pertinent to the rest of its paradigm. Generally speaking, the equality sign denotes "external", mainly tagging, restrictions; the hyphen denotes "internal" restrictions imposed by the lexicon.

The vertical bar "|" denotes tagging ambiguities; the "+" character denotes intrinsic ambiguities (similar to the "external" and "internal" restrictions described above). Both characters follow the entire coding sequence and seperate the two or more tagging alternatives.

Example: Dutch *"jolijt"* would be annotated as Ncms–+Ncns–

NB: The actual characters to be used for tagging restrictions and tag separation were still under discussion at the time of completion of this article. The above description follows the current state of specifications (March 1996) and may be altered in the course of the LE-PAROLE project.

4. REFERENCES

[1]Cloeren (1995): J.Cloeren (ed.): MLAP 93-21 MECOLB: WP5 - Quality Assessment. Tasks 5.2-5.5. Deliverables D11-D14. Final Report. Nijmegen: TOSCA Research Group

[2]Erjavec (1996): T.Erjavec: TELRI WG 5 Report: Corpus Tool Identification *(forthcoming)*

[3]Hladká and Hajič (1995): B. Hladká and J. Hajič: TELRI, Proceedings of the First European Seminar: A Simple Czech and English Probabilistic Tagger: A Comparison. Tihany, Hungary.

[4]Calzolari (1996): N.Calzolari et al. (ed.): MLAP 63-386 PAROLE: WP 4.2.2: Lexicon: Morphosyntactic Specifications: Language Specific Instantiations. Pisa: ILC

[5]Leech and Wilson (1994): G.Leech and A.Wilson: "Morphosyntactic Annotation". EAGLES Document EAG-CSG/IR-T3.1 (Version of October 1994). Pisa: EAGLES Consortium

[6]Ridings (1996): D.Ridings: PAROLE Text Representation. Electronic document on the Göteborg University WWW server. (http://svenska.gu.se/ ridings/textrep/) Göteborg: Sprĺkdata

[7]Volz and Lenz (1996): N.Volz and S.Lenz: MLAP 63-386 PAROLE: WP 4.1.4a: Multilingual Corpus Tagset Specifications (Version of March 1996). Mannheim: IDS

The GRACE action: Grammars and Resources for Analyzers of Corpora and their Evaluation

(Applying the Evaluation Paradigm to Morphosyntactic Ananlyzers for the French Language)

Gilles Adda (LIMSI), Joseph Mariani (LIMSI), Patrick Paroubek (INaLF),
Martin Rajman (ENST)

October 14, 1996

ABSTRACT

Started upon the initiative of Joseph Mariani and Robert Martin, respectively head of the Laboratoire d'Informatique pour la Mecanique et les Sciences de l'Ingenieur (LIMSI) and head of the Institut National de la Langue Française (INaLF), the GRACE action (Grammars and Resources for Analyzers

of Corpora and their Evaluation) aims at applying the evaluation paradigm to morpho-syntactic analyzers for the French language. The first evaluation session: GRACE-I, is devoted to Part-Of-Speech taggers. As a by product of its activities, the action will also make available the language resources it has collected for the evaluation.

After a brief recall on the origins and the nature of the evaluation paradigm, we show how it relates to other national and international initiatives. Then we present the current state of GRACE-I evaluation session relatively to the four components underlying the evaluation paradigm as we see it: corpus building, tagging procedure, lexicon building, evaluation procedure. A presentation of the internal organisation of the GRACE action precedes our conclusion.

I. THE EVALUATION PARADIGM

Recently the evaluation paradigm has been proposed as a mean to foster developments in research and technology for the field of language engineering. Until now, it has been mostly used in the United States in the framework of ARPA project (DoD) on automatic processing of spoken and written language which started in 1994, as well as in the scope of other programs run by the Association for Computational Linguistics (ACL) and by the NIST.

The paradigm is based on a two step process:

- first, create textual or voice data in the form of raw corpora, tagged copora or lexicons, which are then distributed to main actors in the field of language engineering for the realization of natural language processing tools. These tools address problems like disambiguation, natural language database query, message understanding, automatic translation, dictation, oral dialog, character recognition, etc.
- second, the systems are tested on similar data and compared. The results of the test sessions and the discussions ensuing from the publication of the results furnish a sound basis to compare pros and cons of the various methods and systems during a workshop. The resulting synergy is a dynamizing factor for the considered field, here language engineering. For the record, the Linguist Data Consortium whose function is to collect data and organize their distribution is a consequence of programs implementing the evaluation paradigm.

The evaluation paradigm is at the core of the GRACE action, which first focuses on morpho-syntactic analyzers. Later on, it is planned to encompass other aspects of text or vocal data processing. Note that a similar action has been organized in Germany in march 1994 for German morphologycal analyzers (Morpholympics [14] [15]). A following action (Parsolympics) was planned to evaluate parsers.

Another benefit stemming from this sort of project lies in the close collaboration existing between computer scientists and linguists participating in the project. Collaboration is required to define the tag sets, to propose the evaluation criteria, to define evaluation protocols, and to organize the processing of the data. In particular, GRACE tends to promote complementary approaches aiming at the handling on the one hand, of the most frequent cases by the computer scientists and on the other hand, of the rare but difficult cases, by the linguists.

II. RELATED ACTIONS

GRACE is part of the french program: Cognition, Intelligent Communication and Language Engineering("Cognition, Communication Intelligente et Ingenierie des Langues"), shared between the Engineering Sciences department and the Human Sciences department of CNRS (Centre National de la Recherche Scientifique). GRACE has used the scientific contribution of some workgroups of the coordinated resarch group (GDR-PRC) "Communication Homme-Machine" (Speech and Natural Language themes) and has received financial support from the Ministry of Education and Research which is considering to start an evaluation campaign for language engineering products such as spelling checkers. When it will take place, GRACE will constitute an essential complement to it, as it concerns systems working at a lower level of linguistic abstraction.

In one of its programs, the Aupelf-Uref has set up a network for French language engineering (FRANCIL), coordinated by Joseph Mariani. One of the aims of this network is the creation and distribution of language resources for the French language, and the evaluation of natural language processing systems and methods used in language engineering. A large number of laboratories from French speaking countries are contributing to this network. In 1994, the Aupel-Uref has published a series of calls for tenders regarding concerted research actions (ARC-Actions de Recherche Concertees) around the theme of the evaluation paradigm. Two main lines have been identified:

1. linguistics, computer science and written corpora (line A), to address issues in the development of systems for message routing (A1), bi- and multi- lingual corpus alignment (A2) [26], automatic terminology extraction from corpora (A3), and text understanding (A4),

2. linguistics, computer science and oral corpora (line B), with subtopics like dictation (B1) [3], oral dialog (B2), and speech synthesis (B3).

As before, we remark that GRACE articulates harmoniously with these research actions as they concern language engineering tools dealing with a lower level of linguistic abstraction.

In the future, we plan to use the help of Aupelf-Uref to extend GRACE action to other French speaking countries, meanwhile GRACE already collaborates with the SILFIDE national project (Serveur Interactif pour la Langue Francaise, son Identité, sa Diffusion et son Etude) started in 1996 and co-funded by Aupelf-Uref and CNRS. The goal of this project is to organize a network of data servers for language resources for the study of the French language. SILFIDE does not aim at integrating the existing resources (corpora, lexicons and tools) but intends to provide informations on what is available and at which conditions, under a standardized format using the French language as support language. Most of the data held in the SILFIDE servers will be in French, and when it will not be the case, the data concerned will be paired with their equivalent in French. For public domain resources, SILFIDE will offer transfer functionalities. The Centre de Recherche en Informatique de Nancy (CRIN), the Institut National de la Langue Francaise (INaLF), the Laboratoire Parole et Language (LPL), the Groupe d'Étude pour la Traduction Automatique (Geta) and the LIMSI (Laboratoire d'Informatique pour la Mecanique et les Sciences de l'Ingenieur) are participants of SILFIDE.

We contacted Hans Haller of the Institut für Angewandte Informationsforschung (IAI - Saarbr"ucken) for a possible cooperation around the evaluation paradigm between GRACE and the Gesellshaft f"ur Linguistische Daten Verarbeitung (GLDV), which organized the first Morpholympics (evaluation of morpho-syntactic analyzers for German) at Erlangen (Germany) in 1994 [14] [15]. We would like to set-up joint actions, at first involving only laboratories working on French and German, but in the long run, having a European scope for instance in the frame of the ELSNET program (European Network on Language and Speech).

A program, which will include the theme "Linguistic Resources and Evaluation", is in preparation by the European Union. It has a previsionary budget of around 25 Millions of ECUs and lies under the responsability of the language engineering section of the Fourth Framework Program. The call for tender of this program ought to be published near the end of 1996. Such program could rely upon actions undertaken in the scope of LRE programs such as RELATOR for language resources, EAGLES for standardization of language data and evaluation, PP-PAROLE and LE-PAROLE for a network of language resources producers, SQALE for evaluation of dictation systems, or on projects such as MULTEXT, TSNLP or TEMAA (spelling checkers and grammar checkers evaluation [18]). An overview of these LRE projects can be found in [8].

About data re-usability and distribution, we need to mention LINGLINK, a 2 year project also part of the Fourth Framework Program in the sector of Telematics Applications, which will try to promote re-use of resources and tools produced by other projects. LINGLINK will group into domain-related clusters the informations and advertisement means for those project and will organize concerted actions for result and information exchanges. Some of the projects already collaborating with LINGLINK are LE-PAROLE, SPEECHDAT, EUROWORDNET and the ELRA association.

On the issue of re-use and distribution of the data which will be collected for evaluation purposes, GRACE has had initial contacts with Khalid Choukri, the Chief Executive Officer of the European Language Resources Association (ELRA), funded for 3 years by the European Union with the mission to collect, validate and distribute language resources through its commercial counterpart, the European Language resources Distribution Agency (ELDA). A program has been proposed in parallel with the Language Engineering program. It is now being defined by the European Union; we know that it will be called MLIS (*Multi Lingual Information Society*), and that it will have three parts:

- resources and infrastructure
- translation and interpretation
- language engineering

MLIS will be located out of any Research and Development action. One of the goals of the program is to attain a balance between the resources, tools and industries for the various European languages, another is to promote re-use of existiting resources already build by previous projects (e.g. the translation system SYSTRAN). While there will not be any provision especifically made for evaluation actions in this program, means could be set aside to develop specific resources for evaluation, as resources for evaluation is one of the topics identified by the European Unions as topic of interest.

III. THE GRACE ACTION

The project is intended to run over four years (1994-1997). The first year has been devoted to the setting up of a coordination committee and of a reflexion committee. The first phase of the project is devoted to part of speech taggers. A second phase was initially planned to work on syntactic analyzers but has been postponed. In GRACE, we can distinguish four main aspects:

- corpora building
- tagging procedure
- lexicon building
- evaluation procedure

A. CORPORA BUILDING

Generic considerations about corpora for evaluations are mentioned in [23], out of which we quoted some excerpts in the following paragraphs.

Text corpora of sufficient size are required for corpora based systems (see summer ESCA-Elsnet "Corpus Based Methods" in July 1994).

According to the EAGLES report on NLP-system evaluation [10], corpora have proved to be useful mostly in *adequacy evaluation*, and *progress evaluation* leaving aside *diagnostic evaluation* which is the object of the TSNLP project [8].

Generic recommendations on the properties re-usable evaluation data ought to display, are proposed in [9], the report of the study group on evaluation set up within the EAGLES framework, to specify guidelines for assessment and validation of LE projects in the Fourth Framework Program. According to this source, evaluation data must be:

- *realistic*, i.e. be of the same kind as the data received by the system or component being tested during its normal operation,
- *representative*, i.e. contain instances form the full range of input data that would be normally received by the system or component under evaluation,
- *legitimate*, i.e. be easy to acquire, or if not then widely reusable for other purposes

The authors of [9] identify two important properties of the evaluation which strongly condition the requirements put upon test data:

- *the granularity* at which systems will be evaluated (e.g. at the level of user-significant tasks only, or at some level of tasks that are user-transparent),
- *the generality* at which systems will be evaluated (e.g. how much the linguistic features of the provided (resp. expected) input (resp. output) data vary, relatively to the characteristics of the data in the intended application; for instance is the evaluation done against data from different languages, different domains, etc.).

The authors also remark that in most cases, large corpora by themselves do not suffice as the basis for an automatic evaluation procedure. Such corpora need to be annotated depending on the system being tested. But the problem is that most annotation schemes are specific to a particular class of application, and so hardly re-usable. To solve this problem and other difficulties related to evaluation, Klaus Netter has proposed to use, in conjunction with glass-box evaluation,[1] layered annotations for corpora, this at different levels of abstraction matching the intermediate reference levels of the considered evaluation scheme. The annotations could include all kinds of information, such as morpho-syntactic tagging, word sense disambiguation, phrase structures, relational structures, semantic representations including resolved references as well as annonations specific to the application being evaluated.

For GRACE-I, the first evaluation session of the GRACE action, the corpora have two origins, on the one hand, the Frantext database of INaLF, which holds literature texts from the 19th and the 20th centuries amounting to a total of 160 million words and, on the other hand, "Le Monde" newspaper text corpora regularly distributed on a CDROM. Initially 50 millions words were available, but now several hundred millions of words are available with the publication of the archives of the newspaper dating back from 1987.

We have solved the legal issue of copyrights in two ways:

- by selecting from the FRANTEXT database texts which have no copyright restrictions imposed on them.
- obtaining from the "Le Monde" newspaper, through the LIMSI laboratory, the authorization to use and distribute their material, under the condition that each recipient would sign an agreement forbidding redistribution and commercial use of the data it receives.

The data have been separated into three packages corresponding to the first three phases of GRACE-I (for more information, please refer section D), all with a roughly balanced distribution of texts between "Le Monde" and FRANTEXT sources:

- a training corpus of 9 millions word forms,
- a dry-run corpus of 450,000 word forms,
- a test corpus which will have the same size and genre distribution as the dry-run corpus.

The training corpus has been distributed to the participants in January 1996, while the distribution of the dry-run corpus is now in progress, as the participants complete the dry-run phase of the project.

---

[1] [10] defines *glass-box* evaluation as requiring knowledge of the internal working and theoretical underpinning of the system being tested, while black-box evaluation only sees the final output and its relationship to the original input. *Black-box* evaluation is typical of *adequacy evaluation* of market products.

B. TAGGING PROCEDURES.

Tagging procedures deal with three aspects:

- lemmatization
- syntactic categories
- syntactic bracketing

An agreement between experts must also be reached on those aspects. In particular a tagset has to be chosen, along with the means to bracket and mark the text selected to serve as reference material during evaluation. This work concerns both the tagging of text data used for learning and for evaluating the linguistic systems. At first, we were inspired by what has been done at the University of Pennsylvania for the tagging of the Penn TreeBank and upon the tagging recommandations issued by European projects such as MULTEXT and EAGLES. In GRACE-I, the tagset has been defined through a consensus between a reflexion committee composed of a panel of experts and the participants (by means of a circular step-wises refinement procedure, started upon a proposition of the organizers), along with a "tagging manual", a document crafted to help manual taggers in their disambiguation task when preparing the reference material for the evaluation. To illustrate the choices taken and facilitate the discussion, a mini-corpus made of both artificially constructed sentences and excerpts taken from real corpora was hand-tagged and communicated to all the persons concerned before starting the dry-run phase (real size test of the evaluation protocol, with participants implication). To give an idea of how complex it is to preserve consitency over the whole set of documents, the following diagram shows the interdependencies existing between them by means of pointed arrows.

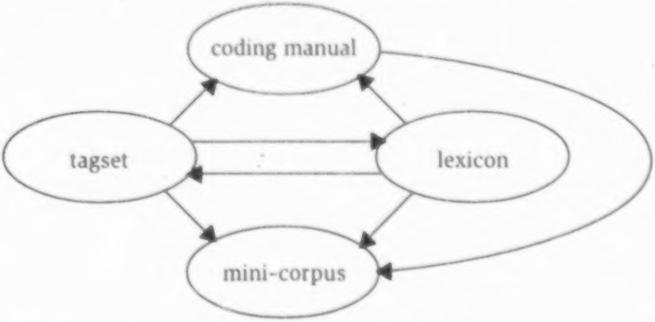


Figure 1: - tagging procedure documents interdependencies -

Note that, in order to build the GRACE tagset, we did not use the technique and tool described in [24], although the spirit of our approach was very similar, because we did not have available at that time a description of the tagsets of all the participants.

C. LEXICON BUILDING

The evaluation paradigm, as we see it, requires lexicons not only to remedy to the possible lack of such resource by some participant, but also to provide training data, to normalize the training corpus and most importantly, to compute some of the error measures (a lexicon provides for every word it contains, the initial level of lexical ambiguity for that word. We are not accounting here for the extra level of ambiguity introduced by possible "category shifts" corresponding to various context specific uses of the word).

For GRACE-I, we studied the availability of existing French lexicons such as INTEXT (LADL) [21], BREFLEX and BDLEX which have both been build in the framework of the "GDR PRC Communication Homme-Machine", lexicons resulting from in-house efforts of the organizers, like for instance the electronic thesaurus extracted from the "Trésor de la Langue Française" (INaLF) and lexicons of the École Nationale Supérieure des Télécommunications (ENST), as well as lexicons resulting from European Unions funded projets like MULTEXT [16]. Note that the goal here is not to create new lexicons from scratch but to see how we can re-use the existing ones, eventually by merging them.

Right now we are using the MULTEXT lexicon [16] as a basis, complemented with a list of compound words extracted from the training and dry-run corpora with the INTEXT toolset, and crossed with INalF TLF thesaurus and the ENST lexicons. This lexicon is called the MULTEXT-GRACE lexicon.

Merging of all the available lexicon data is not yet complete as any change in formalism requires very often a full manual scan of the lexicon to propagate the changes to all the entries. As a result the lexicon has not yet been distributed to the GRACE-I participants.

Some issues are even still debated, among these are:

- the meaning of "non-relevancy", very often for a given linguistic feature and a given word, we associate non-relevancy with contextual undetermination, while for the same cases, a lexicographer will be interested in making the distinction between the two possibilities,
- the coding of prepositions and articles in compound and contracted forms,
- whether to explicit or not category shifts.

D. EVALUATION PROCEDURE

For the determination and the organization of the evaluation sessions, we build upon the work done in previous evaluation actions to guide our reflexion, particularly the evaluation sessions which have been conducted in the United States, especially in the scope of ARPA Human Language Technology action, namely:

- MUC-1, MUC-2, MUC-3 [22], MUC-4 [20], message understanding conferences aiming at the evaluation of message understanding systems, where the systems under test must fill in predetermined forms from texts relating US Navy manoeuvres (MUC-1 and MUC-2) or terrorism acts (MUC-3 and MUC-4),
- TIPSTER, evaluation of automatic information extraction systems from raw text data,
- TREC [12] [13], evaluation of natural language database querying systems,
- ParsEval and SemEval.

MUC and TREC are task oriented black-box evaluation schemes requiring no knowledge of the internal processes or theoretical underpinning of the systems being tested, while ParsEval and SemEval (some of which will be part of MUC-6) are approaches which attempt to evaluate systems at the module level by using a benchmark method based on a reference corpora annotated with a syntactic structure agreed upon by a panel of experts.

We considered also the work of Ezra Black [4][5][6] on syntactic analyzers evaluations done within the scope of an ACL working group, and the "Morpholympics" competition [14][15], which took place in spring 1994 at Erlangen University in Germany and evaluated morphological analyzers for German.

For a list of evaluation methods for lingwares (*linguistic softwares*) now in use in the industry, we suggest to read the report (in French) that Marc Cavazza has written for the Ministry of Education and Research [7]. GRACE is scheduled in four phases:

1. distribution of data for learning (training phase, or *"phase d'entrainement"*),
2. distribution of test data during real size test of the protocole with participants implication (dry-run phase, or *"phase d'essais"*),
3. handling of the evaluations (test phase, or *"phase de tests"*),
4. organization of a workshop for all the participants, where the participants present their methods and their systems and the results of the evaluation are discussed.

In the evaluation procedure initially proposed by Martin Rajman [2] (see next figure), the evaluation is intended to be essentially about the disambiguation power of the systems, distinguishing it from the evaluation of their lexical coverage.

As we cannot impose upon the participants to use the same tagset in their system, we decided to ask the participants to provide us with a mapping table, showing how to project their tags into the GRACE tagset. The translation from one tagset into the common tagset is potentially a many to one correspondence (ideally it is a one to one correspondence).

The measures are done by comparing the input and the output given to the systems. The existence of mapping functions to and from the morphosyntactic description of the reference system (GRACE tagset and lexicon) and the various tagsets used by the different participants authorizes several kinds of ambiguity measures which can be qualified respectively as:

- "absolute" (measure 0 on the figure), i.e. relatively to the reference lexical descriptions themselves;
- "system relative" (measure 1 on the figure), i.e. relatively to the tagset used by the system itself,
- "inter-system relative" (measure 2 on the figure), i.e. relatively to the tagset of another system.

This kind of measure does not take into account the segmentation capability of the taggers, as it depends on concepts very different from the one used in disambiguation algorithms. The quality of a segmenter relies mostly on the amount of information coded in the system (e.g. size of its compound expressions lexicon), while the quality of a tagger is strongly correlated with the algorithm used and the characteritics of the training corpora.

Let us now look a the quantitative aspects of the evaluation of a disambiguation system. For each ambiguous lexical unit, a disambiguation system can :

- perform a "strict" disambiguation, i.e. reduce the list of possible tags to a single tag (which can be valid or erroneous);
- perform a "partial" disambiguization, i.e. reduce the list of possible tags to a smaller list, holding more than one tag, one of which being eventually the correct tag (we say that the system generates a "partial silence");
- do not perform any disambiguation (the system generates a "strict silence").

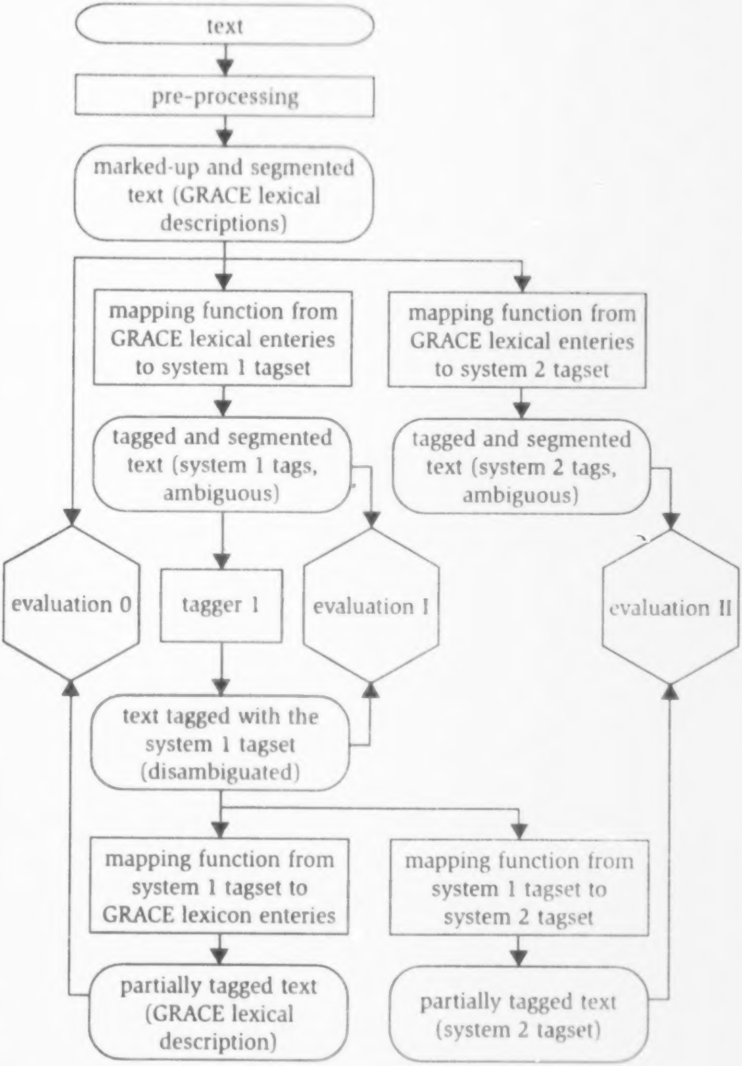

text
pre-processing
marked-up and segmented text (GRACE lexical descriptions)
mapping function from GRACE lexical enteries to system 1 tagset
mapping function from GRACE lexical enteries to system 2 tagset
tagged and segmented text (system 1 tags, ambiguous)
tagged and segmented text (system 2 tags, ambiguous)
evaluation 0
tagger 1
evaluation I
evaluation II
text tagged with the system 1 tagset (disambiguated)
mapping function from system 1 tagset to GRACE lexicon enteries
mapping function from system 1 tagset to system 2 tagset
partially tagged text (GRACE lexical description)
partially tagged text (system 2 tagset)

Figure 2: Proposition de méthodologie d'évaluation.

If we note :

NU	the total number of tokens
NUA	the total number of ambiguous tokens
NUNA	the total number of unambiguous tokens
NDSC	the total number of strict disambiguations which are correct
NDSE	the total number of erroneous disambiguations
NDPC	the total number of partial disambiguations which are correct
NDPE	the total number of partial disambiguations which are erroneous
NE	the total number of disambiguation errors
NSS	the total number of strict silences
NS	the total number of silences (strict or partial)

we can define the following measures:

ambiguity ratio	TA = NUA/NU
apparent disambiguation error ratio	TEa = NE / NU
apparent silence ratio	TSa = NS / NU
apparent correct disambiguation ratio	TDa = (NDSC+NUNA) / NU
real disambiguation error ratio	TEr = NE / NUA
real silence ratio	TSr = NS / NUA
real correct disambiguation ratio	TDr = NDSC / NUA

The evaluation of a disambiguation system could then be caracterized by the triplet: (TA, TEa, TSa), and to compare systems (leaving TA constant), we could measure the distance between the (TEa,TSa) points associated to each system in the plan defined by [0,1]x[0,1].

We may note that the concepts of error and silence defined here are similar respectively to the notions of precision and recall used when evaluating information retrieval systems.

For various practical reasons, it is very unlikely that the measures which will be used in the test phase will be the ones presented above. As the building of the lexicon is still underway, participants have been asked to take raw text (untagged and unsegmented) as input during the dry-run phase; in addition a few participants have been unable to provide a complete mapping table as their system is more a parser than a tagger (the extraction into a legible form, from their system, of the information required to perform the mapping would necessitate a too large effort because of the specificity of the linguistic formalism they use). For the dry-run phase, these participants will perform the mapping of their categories into the GRACE tagset themselves.

The finalization of the definition of the measure function which will be used for the test phase (real evaluation), will be done in concertation with the participants at the end of the dry-run phase.

The problem of the variations of text segmentation between the hand-tagged reference material and the text returned by the participants will be solved by using a re-alignment procedure based on the UNIX command diff. Error acccounting will be turned-off on the portions of texts that the algorithm cannot re-align properly.

IV. ORGANIZATION.

The GRACE coordination committee (in charge of the project management) contains two persons from each laboratory which were at the origin of the project, INaLF and LIMSI and one person from the École Nationale Supérieure des Télécommunications, which joined the project later. It is animated by Patrick Paroubek of INaLF.

The responsability of the reflexion committee is to discuss and to decide which data to make available to the participants, to choose the syntactic categories, to define the evaluation protocol, and to organize the results presentation workshop. It is composed of twenty persons, researchers, computer scientists, linguists, from various laboratories, and is animated by Martin Rajman (ENST) and Gilles Adda (LIMSI).

The third entity of the GRACE organization regroups all the participants to the tests which come both from public institutions or from private industrial corporations. Participant will have to present fully operational systems. Only the participants which have previously agreed to compete in a fair way by providing all the informations required to determine how their system works will be authorized to take part in the workshop concluding the evaluation session.

V. CURRENT STATE OF THE ACTION

The following table provides an overview of the current state of the action. The test phase is scheduled to happen in November '96.

participant	country	mapping table de table provided	mapping table validated	dry-run corpus tagging schedule
ATTBellLabs.	USA			sched.Sept.'96
GREYC(URA1526)	FR	(yes)		done
INGENIA	FR	(yes)		done
CRISTAL	FR	yes		done
IAI	D			
CNET	FR	yes		sched.Sept.'96
RXRC	FR	yes	yes	done
LATL	CH	yes		
LIA/LPL	FR	yes		done
TGID	FR	yes		
ISSCO	CH	yes		sched.Oct.'96
SYNAPSE	FR			
CLIPS	FR	yes		done
ILR/IMS	D	yes		sched.Sept.'96
IBM	FR	yes		sched.Sept.'96
MEMODATA	FR			
GSI-Erli	FR	yes		sched.Sept.'96
CITI	CA	yes		done
INaLF	FR	yes	yes	sched.Sept.'96
ENST	FR			
LIMSI	FR	yes	yes	done

REFERENCES

[1] Abeillé A. et al., "Analyseurs syntaxiques du fran#ais", *T.A. Informations* ISSN 0039-8217, Vol 32-2, 1991.

[2] Gilles Adda, Philippes Blache, Joseph Mariani, Patrick Paroubek, Martin Rajman, "Action GRACE - Mise en place du paradigme d'Evaluation - Application au domaine de l'analyse morpho-syntaxique", proceedings of the Conf#rence sur le Traitement Automatique du Langage Naturel (TALN'95), Marseille, France, juin, 1995.

[3] Frederic Bimbot, "Un point sur les Actions de Recherche Concertees (ARC) - Theme B1", Lettre d'information de l'Aupelf-Uref, N. 3, Avril 1996.

[4] E. Black et al., "A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars", *ARPA HLT Wokshop*, Mars 1991.

[5] E. Black, "Parsing English by Computer: the State-of-the-Art", *International Symposium on Spoken Dialog*, Tokyo, Novembre 1993.

[6] E. Black, "A New Approach to Evaluating Broad-Coverage Parsers/Grammars of English", *Proceedings of the International Conference on New Methods in Language Processing (NEMLAP'94)*, UMIST, Manchester, September 1994.

[7] M. Cavazza, "Méthodes d'évaluation des logiciels incorporant des technologies d'informatique linguistique", *rapport MRE-DIST*, 1994.

[8] R. Cencioni and E. Klein ed., *Linguistic Research & Engineering* (LRE) - *An Overview*, Telematics Programme 1991-1994, Directorate-General XIII, Information Technologies and Industries, and Telecommunications, Commission of the European Communities, June 1994.

[9] Richard Crouch, Robert Gaizauskas, Klaus Netter, "Interim Report of the Study Group on Assessment and Evaluation", EAGLES, Draft report, march 1995.

[10] EAGLES, "Evaluation of Natural Language Processing Systems", *EAGLES report EAG-EWG/PR2*, July 1994.

[11] Fay-Varnier C., Fouqueré C., Prigent G., Zweigenbaum P., "Modules syntaxiques des systèmes d'analyse du français", TSI, Vol. 10-6, 1991

[12] Harman Donna K. (ed.), "The First Text REtrival Conference (TREC-1)", NIST Special publication 500-207, National Institute of Standards and Technology, Gaithersburg MD., 1993.

[13] Harman Donna K. (ed.), "The Second Text REtrival Conference (TREC-2)", NIST Special publication 500-215, National Institute of Standards and Technology, Gaithersburg MD., 1994.

[14] R. Hauser, "Results of the 1. Morpholympics", *LDV-FORUM, vol. 11-1*, Juin 1994, ISSN 0172-9926.

[15] R. Hauser, "The Coordinators' Final Report on the First Morpholympics", *LDV-FORUM, vol. 11-1*, Juin 1994, ISSN 0172-9926.

[16] Nancy Ide, Jean Veronis, "MULTEXT: Multilingual Text Tools and Corpora, Proceedings of the 15th International Conference on Computational Linguistics, COLING'94, Kyoto, Japan, 1994.

[17] G. Leech, A. Wilson "EAGLES Morphosyntactic Annotation, Draft -Work in Progress". Draft technical report, Lancaster, EAG-CSG/IR-T3.1., October 1994.

[18] Sandra Manzi, Maghi King, Shona Douglas, "Working towards User-oriented Evaluation", Proceeding of the Natural Language Processing and Artifical Intelligence Conference, Moncton, N.B., Canada, 1996.

[19] M. Marcus et al., "The Penn TreeBank: Annotating Predicate Argument Structure", *Proceedings of the Conference ARPA'94*, 1994.

[20] MUC-4, *Proceedings of the Fourth Message Understanding Conference*, Morgan Kaufman, 1992.

[21] M. Silberztein, *Dictionnaires électroniques et analyse automatique de textes - Le système INTEX*, Masson, Paris, 1993.

[22] Beth Sundheim, "Third Message Understanding Evaluation and Conference (MUC-3): Phase 1 status report", *Proceedings of the Speech and Natural Language Workshop*, Morgan Kaufman, Pacific Grove, CA, February 1991.

[23] Patrick Paroubek, "Corpora for Evaluation", final report Workpackage 4 Task: 4.1.7, projet PP-PAROLE, LRE 63-368.

[24] Simone Teufel, "A Support Tool for Tagset Mapping", Proceedings of the Workshop SIGDAT (EACL95), 1995.

[25] Jean Veronis et al., "Common Specifications and Notation for Lexicon Encoding", *rapport MULTEXT LRE-62-050*, WP1.6 Deliverable, version pr#liminaire 1994.

[26] Jean Veronis, "Un point sur les Actions de Recherche Concertees (ARC) -Theme A2", Lettre d'information de l'Auplef-Uref, N. 4, Juillet 1996.

Internet addresses The location of the entry WEB page of the GRACE site is:

<URL:http://www.ciril.fr/"pap/grace.html>

The address of the GRACE ftp site is:

<URL:ftp://ftp.ciril.fr/grace>

The Czech National Corpus:
A Brief Survey of the Current State

Prof. PhDr. František Čermák, DrSc.
The Institute of the Czech National Corpus
Faculty of Philosophy, Charles University
nám. J. Palacha 2, Prague 1, 110 00
The Czech Republic
e-mail: Frantisek.Cermak@ff.cuni.cz

The Czech National Corpus (CNC) which is being built up by a concerted effort of a number of academic institutions (mostly universities) is conceived of as a general and possibly representative research source of the contemporary, primarily written Czech Language of the size of some 100 million words in its first stage (while provisions are made for its further growth). Its three other branches include a sample historical corpus, a half-million corpus of authentic spoken language and a general archive serving as a first repository of texts which have been acquired. Out of the many envisageable and possible targets of the CNC the primary one is to serve as a basis for a new dictionary of the Czech language.

At present, some 50 million of textual words might be available in the archive from where, after a clean-up, conversion and TEI/SGML text tagging they gradually flow into the CNC itself. The written part of the Czech National Corpus contains some 30 million words now and the remaining first-stage representative figure of 100 million words might be available in some two years' time. Since the whole project is supported by some government grants where one of the stipulations was to make a substantial part of it publicly accessible, in the spring of this year (1996) first modest version of CNC has gone public. Thus, some 20 millions of newspaper and journal language are to be found now at the following Web WWW pages

http://ucnk.ff.cuni.cz/cnc

with some brief instruction how to search it. Since the respective software management tools are still under development, the access ways are somewhat limited so far, primarily to a concordance form.

The policy pursued so far has been to include whole texts (except advertisements, especially those in English or German, and texts written in Slovak which have been removed); yet this whole-text approach might be questioned later on, as the figures keep growing and some attention might be paid to sampling here, too.

Although a number of minor research problems still have to be solved as far as the envisaged rough picture of the CNC's representative parts is concerned, it was decided to use, as a general background (to be modified), the results of a sociological research undertaken a short time ago. It has been decided to choose, as the sole and primary criterion to be explored in the enquette, that of language reception, i.e. quantitative proportions of various types of language that its users are exposed to, whether actively or in a passiveway. The major results can be summarized in the following (numbers are given in %, the spoken language is not included):

READING	–specialized/technical		33,5
	–nonspecialized		66,5
		–journals	56
		–fiction cum poetry etc	10
		–other	0,5

Just as a marginal information only, let me say that the percentage of the spoken and written language was found to be in the proportion of 67 : 33.

While the proportions of various technical fields can be estimated on the basis of, for example, the circulation, edition and readership of technical journals, primarily, a serious consideration is required in the neglected field of the size of representation of these fields in newspapers and journals of general nature, too. Another related persistent problem not to be found solved anywhere is the double-face quality of many nouns which are both technical terms and general usage words at the same time (bread, pencil etc. are, next to being general, definitely terms, too, at least for their manufacturers) which makes this technical-nontechnical boundary and proportion still more difficult.

Of course, some correction of these figures will be attempted later on, too, stressing other points next to this one, i.e. the language reception.

NANCY TEI Workshop, August 28-31, 1996

Reported by: Mariana Damov, Tomaz Erjavec, Alexander Geyken, Ann Lawson.

■ *MARIANA DAMOV*

e-mail: mariana@ims.uni-stuttgart.de

My project in the LORIA computer pool was to make a prototype for an interractive query system of the TELRI resources using text encoding tools and the DILIB workbench. For this purpose I encoded the list of available in electronically readable form list of TELRI resources into TEI-lite. Following the structure of the document, which was something like a bibliography of data and tools registered by TELRI members, I marked the single items as lists (of resources and tools) and put them together into divisions with the names of institutions as heads. Then I used the DILIB system to recode the file from TEI into SGML, and to index the institution names word by word. As DILIB is a tool compatible with the Web, I designed the query interface for the TELRI resources in HTML format with links to the prepared DILIB routines, so that the queries could be posted and the results shown directly on a Web browser. The described efforts produced a small prototype application of a Web page providing the ability to consult the TELRI resources in an interactive way. The Web site for this prototype is for the time being the TELRI home page in Nancy (http://www.loria.fr/~romary/TELRI/essai.html). It is also currently linked to the TELRI home page in Mannheim (http://www.ids-mannheim.de/telri/whats-new.html) under the item "what's new". I am grateful to Emanuel, Florence and Valeria who spent time to share their experience with me, and assisted me to accomplish my project sucessfully.

■ *TOMAZ ERJAVEC*

e-mail: Tomaz.Erjavec@ijs.si

The main practical task I came to Nancy with was to test-align the English and Slovene version of the novel "1984" by George Orwell, using the XCorpus software. This text is being SGML encoded in the scope of the Copernicus MULTEXT-East project (see http://nl.ijs.si/ME/Corpus/1984/).

In Nancy we managed to sentence segment and align a part of this text, and a demo is available at the Nancy TELRI Web page. However, some problems remain, mostly caused by the heavy markup already present in the current version of "1984", which sometimes confuses the segmenter and aligner. A new version of XCorpus is to be released shortly and will be installed at the Ljubljana site. With this release, we will re-do the alignment and expect better results.

Having watched Ann working with the English text of Plato, I became intrigued and attempted to SGML encode its Slovene translation, which had been produced by ZRC SAZU in Ljubljana. Although I made a good start in Nancy, it took me another two days when I came home to finish this work. In the hope that others (especially members of TELRI WG9) will find the description of this process useful, I made the WWW page http://nl.ijs.si/telri-wg5/Republic/, which describes the up-translation of the Slovene component of the "Republic" corpus.

■ *ALEXANDER GEYKEN*

e-mail: alex@cis.uni-muenchen.de

Apart from the courses we followed in Nancy, I spent the majority of my time in Nancy working on a short extract of a bilingual readings in German and French. My declared goal was not to encode a whole book, but to experience if it is possible in only ONE afternoon via the XCorpus toolbox to encode ASCII text into a SGML/TEI conformant structure, to process sentence alignment on this structure and to display these results on the web. And all this was possible! Of course, the XCorpus tools cannot do any miracles but they are of great help with TEI headers, the hierarchical structure of SGML elements, the correct (re-)numbering of attribute id's and idrefs, and last but not least with sentence alignment.

■ *ANN LAWSON*

e-mail: ann@clg.bham.ac.uk

I spent the majority of the time in Nancy working on the English editions of Plato's "Republic". Having taken both the older (Jowett) and the newer (Harvard) translations to Nancy, I soon realised that the new version was in a very poor state for automatic work. I then concentrated mainly on the old version. I worked on TEI encoding, divison marking, paragraph marking and sentence segmentation. I disentangled various problem cases such as hyphenated words and some quotes, but was unfortunately unable to get a good enough version of the text to align it with another while in Nancy. Hopefully that will soon follow!

New prospective member of the TELRI advisory board

Researches in Central Asia

Hamdam ARZIKULOV,
Laboratory for Language Engineering,
Samarkand State Institute of Foreign Languages
e-mail: hamdam@samarkand.silk.glas.apc.org

The international Speech Statistics Groups held a two-day meeting, on May 20-21, 1996, at the Samarkand State Institute of Foreign Languages - main linguistic university in the new sovereign states of Central Asia where European and Oriental languages are studied. The purpose of the meeting was to explore ways in which research into Turkic language engineering can be integrated to produce a multilingual and polifunctional system so-called Turkic linguistic automaton (TURKLINGTON). The choice of the Samarkand linguistic university is quite understandable: it is known that Samarkand has always been a generally recognized cradle of the Central Asia Moslem culture. Besides the Republic of Uzbekistan is a new Central Asiatic state with the most stable geopolitic and economic situation.

On the other hand the independence status of the new Turkic republics necessitates to create their own information industry. Therefore language engineering (LE) turns with interest to Kazakh, Kyrgyz, Uzbek, Azerbaijan and other Turkic languages. In Kazakhstan the research body headed by Prof. K.Bektayev is creating a machine fund (i. e. thesaurus) of the Kazakh language, working on English-Kazakh MT and statistical-informational typology of Turkic texts. The Uzbekistan LE research group headed by Prof. H. Arzikulov consists of three teams (Samarkand, Tashkent and Nukus collectives). The group is engaged in creating a machine thesaurus of Uzbek and Karakalpak languages, in designing English-Uzbek and Uzbek-English MT systems and computer-aided language learning (CALL) of the Uzbek, English and French languages. In addition, the Samarkand research team works out MT patterns for Arab and Persian languages. In Bishkek (capital of Kirgizistan) Prof. T. Sadykov and his colleagues from the Kyrgyz Academy of Sciences are developing methods of automatic Turkic text synthesis. Ph.D. M.Aiymbetov from the Karakalpak Pedagogical Institute studies statistical proprieties of Turkic texts. The models of formal morphological analysis of Turkic word-forms are worked out by Prof. M. Mahmudov's LE group in the Azerbaijan Academy of Sciences in Baku.

The meeting at the Samarkand State Institute of Foreign Languages was attended by 52 people from Turkic and Russian academic, university and industrial sities. It was opened by Prof. Yusuf Abdullaev, the rector of Samarkand State Institute of Foreign Languages (SSI of FL), followed by an introduction to the synergetic problems of NLP by acad. R.Piotrowski (Hertzen Univ of Russia). The second talk was given by Prof. T. Sadykov (Kyrgyz Academy of Sciences) who described the state of the art in automatic analysis of Turkic word-forms and its morphological aspects.

Prof. H.Arzikulov (SSI of FL) pointed out that commercially viable NLP - systems depend crucially on getting access to real text patterns. Therefore automatic dictionaries and machine grammars are developed in the TURKLINGTON not on the well known dichotomy "Language - speech" but on the basis of the trichotomy "Language system - speech system - text". It is important to emphasize that NLP leaned upon the language system produces a primitive lexico-grammatical translation, where as a linguistic automaton working with speech pattern it would provide a more adequate MT, text abstracting or spell-checking. Asst.Prof. M. Aiymbetov (Nukus Univ) presented a new taxonomy classification of Turkic languages and dialects on the basis of their lexico-statistical properties.

After thePlenary Session three Section sessions were organised. The first session, "Computer-Aided Text Processing", began with the talks of Asst.Prof. D.Urinbaeva (Samarkand Univ) entitled "Automatic Analysis of Amir Timur's Works" and that of Prof. B. Urinbaev (SSI of FL) "Lexico-Grammatical Features of Amir Timur's Works". The third talk "Computer-Aided modeling of mathematical terminology of the Tamerlan's epoch" was given by Asst. Prof. I.Hojiev (SSI of FL). Then Prof. B.Tursunov and M.Begmatov (SSI of FL) tried to convince the audience about the importance of for mal specification of textual unities for automatic pattern recognition. A survey of Turkic text automatic analysis was presented by Asst.Prof. Garipov (Bashkyriya Univ) and Prof.R.Kilichev (SSI of FL).

The second session considered "Computer-Aided West and Oriental Language Learning" and was opened by S. Doniyarova (SSI of FL), who talked about "Semantic Field in Lexics and its Computer Application in Language Learning". The second talk - "Teaching Computer-Aided Grammar" - was given by U. Umirzakov (Samarkand Univ). M.Choriyev (Karshi University), M. Boliev (Samarkand Med.School) and I.Akramova were the main participants in this discussion.

The third session of the meeting focused on the topic of "Computer Programs for NLP". This session was opened by Asst.Prof. A.Karshiyev, who gave a very interesting talk on Machine Translation from English into Uzbek. The talk "Design and implementation of a spell Check for Uzbek Language" was presented by E.Gujov (Samarkand Univ). This was followed by a series of short presentations on Automatic Analysis and Synthesis of Natural Language (U.Urinbaev, Samarkand Univ; O.Kholmurodov, Ssi of FL; S.Kobylov, Samarkand Univ; M.Ayimbetov, Nukus Univ).

Some Interesting Events – Past and Future

ESSLLI'96

Geert-Jan M. Kruijff , chairman of ESSLLI '96
e-mail: gj@ufal.mff.cuni.cz

After summerschools in Groningen (1989), Saarbrücken (1990), Leuven (1991), Colchester (1992), Lisbon (1993), Copenhagen (1994), and Barcelona (1995), this year's summerschool was held in Prague, Czech Republic, from August 12 until August 23, 1996. Alike the other summerschools, the main focus was the interface between logic, linguistics, and computation, particularly where it concerns the modelling of human linguistic and cognitive abilities. As such, the programme included courses, workshops and symposia covering a variety of topics within six areas of interest: Logic, Language, Computation, Logic and Computation, Computation and Language, and Language and Logic. Examples were Ivan Sag's symposium on "Syntax and Semantics of Coordination" (Language), Patrick Cousot's introductory course on "Abstract Interpretation" (Computation), John Carroll's workshop on "Robust Parsing" (Language & Computation), and Johan van Benthem's advanced course "Dynamic Logic and Information Flow".

Besides the approximately 55 courses, given by -in total- 70 lecturers from all over the world, there were also three invited evening lectures. This year's lectures were given by Barbara Partee ("Quantificational Domains, Focus, and Recursive Contexts"), Petr Sgall ("Prague School through the Epochs"), and Johan van Benthem ("The Common Concerns of Logic and Philosophy of Science").

ESSLLI'96 was attended by more than 440 people (including lecturers), among them being 40 grantees (in part funded by the Volkswagen Stiftung and the International Institute of the University of Tübingen/EACL). Except for Antarctica, all the world's continents were represented - making ESSLLI into more than just a (major) European experience!

Next year, ESSLLI will be held in Aix-en-Provence, France. For more information on ESSLLI'97, please go to their website at

http:// www.lpl.univ-aix.fr-/ -esslli97

or send an email to

esslli97@lpl.univ-aix.fr

ESSLLI'96 was organized under auspices of FoLLI (the European Association for Logic, Language and Information), Charles University, and the Czech National Technical University (ČVUT).

Vilém Mathesius Lectures Series 10

Prague, February 10-21 1997
Organized by the Institute of Formal and Applied Linguistics,
Charles University, Prague
Hotel Krystal, Prague 6, José Martí Street

- Emmon Bach (Univ. of Massachussetts, Amherst, USA):
  Varieties of polysynthesis
- Elisabeth Engdahl (Göteborg, Sweden):
  Recent developments in theoretical syntax
- Fred Jelinek (Johns Hopkins University, baltimore, USA):
  Stochastic methods in linguistics
- Ferenc Kiefer (Budapest, Hungary):
  The morphology syntax interface
- Bente Maegaard (Copenhagen, Denmark):
  Evaluation of natural language processing products
- Peter W. Nesselroth (Canada):
  What is the deconstruction and why are they saying such terrible things about it
- Ellen Prince (Univ. of Pensylvania, Philadelphia, USA):
  Syntax-discourse interface
- Helmut Schnelle (Univ. of Bochum, Germany):
  The structure of language and the topography of language areas in the brain
- John Sinclair (Birmingham, Great Britain):
  Computerized lexica
- Oliviero Stock (Trento, Italy):
  Chart parsing and bidirectionality
- Eloise Jelinek (USA):
  Some specific phenomena of non-Indoeuropean langauges

The following Czech professors have also been invited to give a talk:

- Jan Hajič:
  Computerized corpus of Czech language
- Eva Hajičová:
  Recent research in topic-focus articulation
- Oldřich Leška:
  Czech structural linguistics
- Jarmila Panevová:
  Dependency syntax
- Jaroslav Peregrin:
  Some issues of theoretical semantics
- Petr Sgall:
  Typology

The number of participants is limited. There are grants available for students from the post-communist countries covering the tution fee (including accommodation); applications for grants must be submitted before October 31st, 1996, with a recommendation of the student´s supervisor (professor or senior researcher). The tution fee (including accommodation in double rooms with private showers and with buffet breakfast for 13 nights, lunches for 10 weekdays, a welcome party and all teaching materials) is 380 USD.

Further information:

{brdickov, hajicova}@ufal.ms.mff.cuni.cz

List of Participants[*]:

- ANDERSEN Poul
  e-mail: m764@eurokom.ie
- BECI Bahri
  NEW!!!
  e-mail: beci@igjl.tirana.al
- BENKO Vladimír
  e-mail: jazybenk@savba.savba.sk
- BIEN Janusz S.
  e-mail: jsbien@plearn.edu.pl
- ČERMÁK František
  e-mail: frantisek.cermak@ff.cuni.cz
- ERJAVEC Tomaž
  e-mail: et@cogsci.ed.ac.uk
- FISIAK Jacek
  e-mail: fisiak.plpuam11.bitnet
- GELLERSTAM Martin
  e-mail: gellerstam@svenska.gu.se
- HAJIČOVÁ Eva
  HLADKÁ Barbora
  e-mail: hajicova@ufal.mff.cuni.cz
  hladka@ufal.mff.cuni.cz
- JAKOPIN Primoz
  e-mail: primoz.jakopin@uni-lj.si
- JAROŠOVÁ Alexandra
  e-mail: sasaj@juls.savba.sk

NEW!!!

- LAURENT Romary
  e-mail: Laurent.Romary@loria.fr
- KRUYT Truus
  e-mail: kruyt@rulxho.leidenuniv.nl
- MARCINKEVIČIENÉ Rúta
  e-mail: ruta.marcinkeviciene@vdu.lt
- OIM Haldur
  e-mail: hoim@psych.ut.ee
- PAJZS Júlia
  e-mail: pajzs@nytud.hu
- PASKALEVA Elena
  e-mail: hellen@bgearn.bitnet
- PENCHEV Iordan
  e-mail: jpen@bgearn.bitnet
- SINCLAIR John M.
  e-mail: j.sinclair@bham.ac.uk
- SPEKTORS Andrejs
  e-mail: aspekt@mii.lu.lv
- TEUBERT Wolfgang
  VOLZ Norbert
  e-mail: telri@ids-mannheim.de
- TUFIS Dan
  e-mail: tufis@roearn.ici.ro
- ZAMPOLLI Antonio
  e-mail: paula@icnucevm.cnuce.cnr.it

---

[*] You can see detailed addresses in Newsletter No. 2.

WHAT IS TELRI

The Concerted Action TRANS-EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE (TELRI), is a COPERNICUS project funded by the European Commission. TELRI has a duration of three years (1995-1997). It brings together 22 institutions of 17 European countries (Albania, Germany, Great Britain, Slovakia, Italy, Bulgaria, the Czech Republic, Sweden, Slovenia, Romania, Estonia, France, the Netherlands, Latvia, Lithuania, Poland and Hungary).

TELRI is setting up a permanent network of leading national language and language technology centres in the whole of Europe. It pools existing language resources, corpora, machine-readable dictionaries and lexicons, lexical databases, and generic software tools for the creation, re-use, maintanance, validation , and exploitation of linguistic data. It complements these repositories with newly created multilingual resources, offering a wide range of language data to the NLP community. TELRI is establishing a platform where research and industry meet, exchange resources and engage in product-oriented cooperation.

Links have been established with language centres elsewhere in Europe, with relevant European organizations and ventures, and with focal language institutions in other parts of the world.

FOR INFORMATION.

Inquiries about TELRI may be addressed to: **Dr. Wolfgang Teubert,** Institut für deutsche Sprache, P. O. Box: 101621, 68016 Mannheim, Germany, Phone: +49 621 1581 437, Fax: +49 621 1581 415, e-mail: telri@ids-mannheim.de

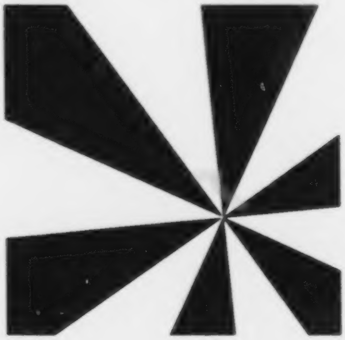
TELRI's WWW Document.

Detailed information about TELRI and its activities is available through the World Wide Web (WWW) at the following URL: **http://www.ids-mannheim.de/telri/telri.html**
Webmaster: Alena Böhmová, e-mail: webadm@smetana.ms.mff.cuni.cz
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics,
Charles University, Malostranské nám. 25, 118 00 Prague 1, Czech Republic

TELRI

TRANS EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE

Concerted Action in the Framework of the Copernicus Program



Newsletter 5 *April 1997*

Contents

1. Editorial --------------------------- 3
2. TELRI Event - Ljubjlana workshop ------------ 5
3. News from TELRI Working groups ----------- 28
4. On the Lexicons ------------------------ 30
5. Some interesting events - past and future ------- 36
6. List of Participants ---------------------- 39

Coordinator:

Dr. Wolfgang Teubert
Institut für deutsche Sprache
P. O. Box 101621
D - 68016 Mannheim, Germany
phone: +49 621 1581 437
fax: +49 621 1581 415
e-mail: telri@ids-mannheim.de

Editors:

Prof. Eva Hajičová
Mgr. Barbora Hladká
Institute of Applied and Formal
Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25
118 00 Prague 1, Czech Republic
tel.: +420 2 21 91 42 57
fax: +420 2 21 91 43 09
e-mail:
hajicova@ufal.mff.cuni.cz
hladka@ufal.mff.cuni.cz

Contributions to TELRI newsletter, and address corrections, should be sent to:

e-mail:
hladka@ufal.mff.cuni.cz
fax: +420 2 21 91 43 09

Editorial

Ann Lawson, *University of Birmingham*

The Working Group "Joint Research" is amongst the most fruitful research projects within the framework of the TELRI project, offering partners spanning most of Europe the opportunity to work together on a close and practical level. This group set out to collect, encode and distribute electronic text versions of a sample text, Plato's "Republic", with the aims of testing corpus alignment software and investigating new methods of extracting data about translation equivalents, collocations and phraseological units in various languages.

This newsletter provides a sample of the work currently ongoing in the encoding and alignment of parallel language texts, and the research subsequently carried out. The papers were presented at a weekend workshop on the 1st and 2nd February 1997 in Ljubljana, at the Jozef Stefan Institute, Slovenia, in which many members of Working Group "Joint Research" participated. The workshop was well timed, as most partners present had not only made an electronic version of Plato's "Republic" available, often encoded with additional information, but had also undertaken some preliminary work on parallel texts. The weekend offered a wonderful opportunity to discover first-hand the tasks other partners are tackling, the problems and issues they are encountering and perhaps most importantly of all, the solutions found to deal with them successfully.

In all, around 20 TELRI partners participated in the workshop, with local students and visitors swelling the number. Tomaz Erjavec organised the practicalities of weekend admirably, including computer access, accommodation and entertainment.

The weekend was split into four distinct sections according to the tasks involved and the papers being presented.

The first session dealt with the encoding of electronic texts in Standard Generalised Markup Language (SGML), sentence segmentation and structural markup in general. An afternoon was then spent looking in detail at the issue of alignment, in order to analyse parallel texts. The ideal, theoretical and practical considerations were explored with the benefit of first-hand experience and computer demonstrations.

The third session looked at partners' experiences and results of comparative analyses of translations, and their bearing on future research. The issue of translation equivalents in the texts as compared to traditional bilingual dictionaries was of special interest to many partners.

The final afternoon was spent discussing plans for the future. The Workshop and this publication have fuelled the motivation and enthusiasm of the partners to continue this line of work. All left the workshop with ideas and plans for both the near future in terms of further segmentation, alignment and analysis of texts, and for the longer-term future.

The Third TELRI Seminar, planned for late October 1997 at the Tuscan Word Centre, will discuss in more detail the issue of "Translation Equivalence". It will be an opportunity for TELRI to present further research findings to a wider audience of industry representatives. The Ljubljana workshop proved to be a vital intermediate stage towards such future ambitions.

More info on the Workshop is at *http://nl.ijs.si/telri-wg5/LjWS.html.*

TELRI Event – Ljubjlana workshop

A Sample Analysis of Ways of Expressing AGREEMENT and DISAGREEMENT in the Translation of Plato's Republic into Czech and English.

František Čermák
Institute of Czech National Corpus
Charles University,
Prague, Czech Republic
e-mail: Frantisek.Cermak@ff.cuni.cz

0. PLATO'S TRANSLATED TEXTS IN CZECH AND ENGLISH

The paper Czech version (Ústava) used has been that of the second Czech edition published by Svoboda/Libertas in Prague in 1993 and it is based on the Greek original as published in Paris 1932 by E. Chambry (Platon, Oevres Completes, tome VI, La république I-III. The first Czech translation is that of R. Hošek (the first translation being from J. Novotný, published in 1921, both published in Prague).

The paper English text used has been that of D. Lee, published in Penguin Books in its 2nd revised edition in 1974.

To get an electronic version, the Czech text had to be scanned, revised and corrected in various laborious ways and stripped of any additional text added by the editors.

1. HANDLING OF BOTH TEXTS.

At the time of the analysis there has been, unfortunately, neither a suitable aligner nor the English tagged text which I planned to use in a comparative study of Czech and English translation of Plato's Republic. The Czech version exists now in two forms: as a plain text and in the SGML format. My aim has been to analyse, at this stage, a sample of both texts as to the variability of corresponding expressions of agreement and disagreement. The Czech text, which has been tagged by the side numbers-cum-letters (a-b-c-d-e) but which could not use these markings as there has not been a corresponding English text at the time, however, has been hampered in its analysis in that it has then had to be analysed manually, together with the English one. This was due to a suitable aligner being unavailable. I have to say that the analysis and its preparation took, unfortunately, much more time than planned as both texts differ considerably in their treatment and distribution of text paragraphs which simply cannot be relied on. This must be somewhat of a disappointment for the alignment philosophy which so heavily relies just on paragraph alignment primarily.

2. *AN ANALYSIS.*

The random choice of the sample to be analysed fell on Chapter 2 (The Individual, The State, and Education). In the subsequent step, next to checking the overall distribution of both positive and negative expression of the attitude (2.1), I decided to take up, in some detail, the negative attitude, i.e. disagreement (2.2).

2.1 Overall Distribution of both Agreement and Disagreement:

CZECH:	Agreement	102x :	
	-direct (yes/yes (ano) + qualifier):		24
	-indirect (other than "yes", much longer and diverse):		78
	Disagreement	23x :	
	-direct (no/no (ne) + qualifier):		22
	-indirect (other than "no/not"):		1
	Disagreement not expressed negatively (as against English) or found missing		5
ENGLISH:	Agreement	110x :	
	-direct (yes/yes + qualifier):		18
	-indirect (other than "yes", much longer and diverse):		92
	Disagreement	20x :	
	-direct (single no/not/im- or with qualifier):		20
	-indirect (other than "no/not"):		0
	Disagreement not expressed negatively (as against Czech) or found missing		8

No attempt has been made, at this stage, to go into what is termed here "qualifier", but some illustration of what is meant can be seen from the examples below.

Some Conclusions:

A-Agreement : Disagreement in both languages approximately = 5 : 1

B-Agreement in both languages: 4 times more indirect than direct, while

C-Disagreement has a reverse situation: 19 times higher preference for a direct disagreement than for an indirect one. Moreover: diasagreement seems to rely, more than agreement, on various verb forms.

2.2 Distribution and Analysis of Disagreement Forms:

Only very general formal features have been picked up and scrutinized, such as presence or absence of morphological negation and cases of no correspondence. Numbers are auxiliary page numbers of the printouts used for the analysis.

ENGLISH:		CZECH:
Nonsense!		Nepovídej!
There can be no other		Nevidím
No doubt	x	xOvšem
Impossible-repeated		To opravdu možné není
Not at all		Nikoliv
xI dare say	x	Já to nejsem schopen rozpoznat
No		Ne
xQuite true	x	Vůbec ne
No doubt	x	xI já si to myslím
xCertainly	x	Jak by ne?
xWe must	x	To opravdu ne
Certainly not		Nezdá se
xMost assuredly	x	Nemůže to být jinak
Certainly not		To tedy ne
We cannot		Ne, to v žádném případě nedovolíme
-	x	Při Diovi, ne!
No, indeed		Vůbec ne
Certainly not		Ani to ne
No		-
Impossible	x	xJak by bylo?
I cannot answer you		To teď ovšem nemohu takto tvrdit
He cannot		Jak by ne?
Impossible		To je nemožné
Heaven forbid		Ovšem, to ne
I cannot say		To nevím

There is nothing more hateful to them	x	Ovšem, velmi
That is inconceivable		Ani zdaleka
None whatever		Neexistuje!

2.3 A Commentary

It is certainly alck of basic, i. e."negative", correspondence here which is most surprising. It boils down to two cases:

1- The translator has opted for the opposite polarity of expression (8 times in English and 5 times in Czech) which amounts to some astonishing 20 or more percent. One can only wonder what the Greek original has in these cases. Or

2- There is simply no corresponding form used at all, either positive or negative in one of the languages involved (one case in each language).

An interesting matter for conversation analysis and that of politeness is the obvious preference here, contrary to positive, i. e. agreement expressions, to shy at simple negative one-word expression. Also ways how these negative expressions are expanded is of much interest. To be able to arrive at any typicality here and what is less typical, one would need a much larger text, of course, and this is what the subsequent stage of the comparative analysis should be concerned with.

3. CONCLUSION AND A SUGGESTION

I think that an analysis like this, if expanded, might lead to a number of insights and could fit in into a broader mosaic of both text analysis and search of translation equivalents, even though the latter is somewhat difficult once the original language is not used.

However, it is the basis of a reliable alignment which has to be found, which, as my experience shows, is not to be sought in paragraphs but in the broader a-b-c-d-e notation, a conclusion which may not be general. Unfortunately, any more detailed mark-up requires an enourmous amount of work and checking.

My thanks go to my colleague, Renata Blatná, who has helped with the working annotation of text and selection of pertinent passages.

Slovene Translation: Structure Markup in TEILite

Tomaž Erjavec
Language and Speech Group, Intelligent Systems Dept.,
Jozef Stefan Institute,
Ljubljana, Slovenia
e-mail: Tomaz.Erjavec@ijs.si

1. INTRODUCTION

The Slovene translation of Plato's 'Republic' was keyed-in by the Ljubljana 2 site (Institute for Slovene Language "Fran Ramovš", Slovene Academy for Sciences and Arts, Ljubljana, Slovenia) in the text editor Eva. This version served as the starting point for encoding the document in TEI Lite conformant markup. The task of the uptranslation was begun at the summer Nancy workshop, and finished in Ljubljana. In total, the process took about three days. The up-translation was accomplished partly via search and replace operations and macros in the editor Emacs, and partially via small Perl programs.

TEI Lite (*http://www.uic.edu/orgs/tei/lite/*) is a simplified version of the Text Encoding Initiative (TEI) Guidelines, which are addressed to anyone who wants to interchange information stored in an electronic form. As explained in the TEI Lite introduction, the TEI Guidelines provide a means of making explicit certain features of a text in such a way as to aid the processing of that text by computer programs running on different machines. This process of making explicit we call markup or encoding. Any textual representation on a computer uses some form of markup; the TEI came into being partly because of the enormous variety of mutually incomprehensible encoding schemes currently besetting scholarship, and partly because of the expanding range of scholarly uses now being identified for texts in electronic form. The TEI Guidelines use the Standard Generalized Markup Language (SGML) to define their encoding scheme. SGML is an international standard (ISO 8879), used increasingly throughout the information processing industries, which makes possible a formal definition of an encoding scheme, in terms of elements and attributes, and rules governing their appearance within a text. In selecting from the several hundred SGML elements defined by the full TEI scheme, a useful 'starter set' has been identified comprising the elements which almost every user should know about. Experience working with TEI Lite is invaluable in understanding the full TEI and in knowing which optional parts of the full TEI are necessary for work with particular types of text.

The simplicity, availability and extendibility of TEI Lite were the principal reasons why it was chosen as the markup scheme for the Slovene Plato. In the rest of this paper we give the structure of the TEI Lite marked-up Slovene Plato and some possible directions for further work.

2. STRUCTURE OF THE CORPUS

As regards character representation, all non-ASCII characters appearing in The Slovene Plato (the corpus) have been substituted by SGML entities; the following entities have been used:

- &Ccaron; &ccaron; &Scaron; &scaron; &Zcaron; &zcaron;
- &aacute; &eacute; &iacute; &oacute;
- when hyphen was used for intra-sentential punctuation, it was substituted by &mdash;

The corpus consists, as every TEI (Lite) document, of a header and a body. The header has four major divisions which together provide a detailed documentation of:

1. the electronic document itself and the sources from which it was derived;
2. the encoding system which has been applied;
3. descriptive information categorizing the document and its subject matter;
4. its revision history.

We will not further discuss the Slovene Plato corpus header here; suffice it to say that it has 220 lines and 105 elements detailing the above four categories.

The body of the text uses further markup. In particular, the following elements are distinguished in the body of our corpus, given together with the number of times they occur:

- DIV (203)
- HEAD (203)
- P (4252)
- LG (46)
- L (113)
- Q (4549)
- XPTR (598)

The DIV elements encodes textual divisions; we made use of two levels of DIV. The top level divides the Slovene Republic into 10 books and is marked by <DIV type="part">. Each such <DIV> is followed by the header (HEAD) of the part, which contains the text "BOOK ONE" etc. In the original, certain paragraphs are preceded by a number; these numbers were taken as second level divisions, and encoded as <DIV type="section">. Each such DIV is followed by its HEAD, containing the text "1." etc.

Paragraphs are marked by P elements, while poems or fragments of poems have been marked up by the LG (line group) elements, with each line of the poem marked by L.

The speeches of the participants of the dialogues are marked by Q elements. The opening and closing quote marks have not been in these cases retained in the corpus. Furthermore, Q elements are used to denote 'mentioned' words, so Qs also appear within Qs. For the 'mentioned' cases, quote marks have been preserved. This gives us a structure as e.g. "He said: <Q>The word is <Q>'honour'</Q></Q>." However, in the dialogues a speech often spans several paragraphs. As TEI Lite does not allow Q elements to encompass P elements, such Qs have been split to be P internal.

Finally, the text makes reference to endnotes; these have been marked up as external pointers (XPTR), e.g. with endnote number 9 being marked up as <XPTR N=endnote_number>. The XPTR element has been inserted at the closest legal TEI Lite position to where the original endnote number appeared.

3. FURTHER WORK

The TEI Lite Slovene Plato has been structurally marked-up and is available from the IDS ftp server; a WWW document describing this corpus and giving samplers of the original EVA Plato, the TEI Lite version and its HTML rendering is available via *http://nl.ijs.si/telri-wg5/Republic/*

For translation studies, alignment to at least sentence level of the various translations is essential. For this it would be advisable for all the Plato translations to be first TEI encoded as the Slovene was, and then sentence segmented. With such a structure, it is then relatively easy to sentence align the various translations, using one of the available implementations of the the Church & Gale algorithm, as has been done e.g. in the MULTEXT-East project (*http://nl.ijs.si/ME/*).

Part-of-Speech Tagging in the Slovenian Translation of Plato's Republic

Primož Jakopin, Aleksandra Bizjak
Institute for the Slovenian Language Fran Ramovš,
Scientific Research Centre of the Slovenian Academy of Sciences and Arts,
Ljubljana, Slovenia
e-mail: primoz.jakopin@uni-lj.si, aleks@zrc-sazu.si
Internet homepage: http://www.zrc-sazu.si/isjfr/telri/platon.html

In the fall of 1996 the problem of part-of-speech tagging for Slovenian was addressed as a part of the preparatory work for the upcoming new lexicons. As such tagging for Slovenian has not been approached before, the most relevant reference is the work conducted in the frame of the Copernicus project MULTEXT-East. At the time of writing this contribution an up-to-date report of this project can be found on: *http://nl.ijs.si/ME/Lexica/MorphSyn/mte-D11M/mte-Da11M.html* (Work Package WP1 - Task 1.1, Deliverable D1.1, Version 2.2, Milestone M, Intermediate Report, 16 October 1996). The MULTEXT-East project is coordinated by CNRS from Aix-en-Provence, France, and the languages of participants are Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovenian. The tagset developed for the project is very comprehensive and encompasses all the features of the languages mentioned. However, for the purposes of any large-scale tagging effort for only one language (Slovenian) and with lots of human-machine interaction, it was assessed as inappropriate. It has been designed to be complete and universal, and it is therefore more suitable for the software use. Its main drawback from our point of view is that the tags themselves are either not readily decipherable by the linguist performing tagging or verification of automatically tagged texts, or are so long that the tagged text has to be displayed word per line.

Let us illuminate this point with two sentences. The first, *Seveda se lahko motijo.* 'Certainly they may be wrong.', tagged according to MULTEXT-East:

Seveda	Particle
se	Pronoun reflexive accusative yes personal nominal
lahko	Adverb general positive
motijo.	Verb main indicative present third plural no

and according to the tagset used in the project:

Seveda	*se*	*lahko*	*motijo.*
Č	Gmp	A	Gcp

where Gmp stands for *separate verbial morpheme* and Gcp for *verb, third person, plural.*

The second sentence below, using the proposed tagging method, fits in two lines what would otherwise be more like a full screen of data:

In	*s*	*tem*	*trdiva*	*ravno*	*nasprotno*	*od*	*tega*	*kar*	*pravi*	*Simonid.*
Vpr	E6	ZSKse6	Gce	A	Sse4	E2	ZSKse2	ZVR	Gce	IOmel

Its English translation, cut out of the wider context: *...if so, we shall be saying the very opposite of that which we affirmed to be the meaning of Simonides.* The sentence and its translation also show clearly the vast range of the problem of translation equivalents.

The novel *Pomladni dan* 'A day in spring' by Ciril Kosmač, a post-war Slovenian writer, known for his excellent style, has been chosen as the testing ground (179 pages, 61.532 words). In the course of performing the task the tagset has evolved and the required tagging facilities have gradually been added to EVA, the lingware-oriented editor developed by P. Jakopin.

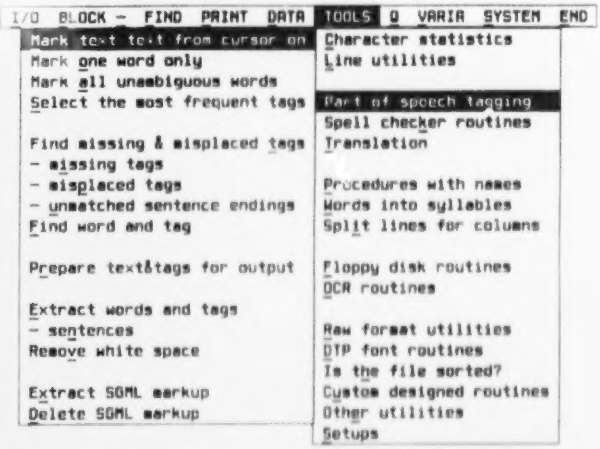

I/O BLOCK - FIND PRINT DATA TOOLS Q VARIA SYSTEM END
Mark text text from cursor on
Mark one word only
Mark all unambiguous words
Select the most frequent tags
Find missing & misplaced tags
- missing tags
- misplaced tags
- unmatched sentence endings
Find word and tag
Prepare text&tags for output
Extract words and tags
- sentences
Remove white space
Extract SGML markup
Delete SGML markup
Character statistics
Line utilities
Part of speech tagging
Spell checker routines
Translation
Procedures with names
Words into syllables
Split lines for columns
Floppy disk routines
OCR routines
Raw format utilities
DTP font routines
Is the file sorted?
Custom designed routines
Other utilities
Setups

Figure 1: EVA tagger submenu with the path leading to it

The "TOOLS" main menu entry has obtained a new topic, "Part-of-speech tagging" with 15 functions; the last two are dedicated to SGML markup. The tagger makes use of four files, which reside in memory:

1. the text itself, where every second line is blank to allow the assignment of tags;
2. the dictionary of words and their tags with frequencies, which expands on the fly;
3. the explanations for unpacking of any tag into its full description (i.e. IOme1 into name, personal, masculine, singular, case 1);
4. the history file, which is used for disambiguation.

All the files are standard EVA editor files, which makes any changes or corrections quite straightforward. The tagging is not fully automatic yet (disambiguation is still not fully operational); it could be described as computer-assisted.

So far *Book 1* of Plato's Republic has been tagged and verified (9542 words out of 92730, i. e. a little over 10%). There are 574 different tags (489 in the first 9542 words of *A day in spring*, 1003 in all). The part-of-speech distribution for both texts shows a higher share of pronouns, conjunctions and particles with considerably less verbs in Plato's text compared to a standard Slovenian novel. In total there are 1.53% of names (added to nouns) in the Book 1 and 1.93% in *A day in spring*. The sample of *Book 1* is probably too small for any analysis but one relation is interesting enough to deserve mention - the gender of nouns. It is 46.5% (m), 39.1% (f) and 14.4% (n) for the Plato's text and 40.3% (m), 47.1% (f) and 12.6% (n) for *A day in spring*. The names have contributed largely to the high frequency of masculines, without them the relation would be 41.5% (m) versus 42.7% (f). There are 145 male names in *Book 1* and a single female one (Paeania), which even fails to show up in the English translation (*from Paeania* in Slovenian, *Paeanian* in English).

Part of speech	Book 1	A day in spring
verbs	25.24 %	32.94 %
nouns	18.00	18.45
pronouns	16.50	11.47
conjunctions	12.48	8.76
prepositions	7.42	8.55
adjectives	6.78	7.02
adverbs	6.25	5.99
particles	6.96	5.71
numerals	0.31	0.64
interjections	0.05	0.44
abbreviations	0.01 %	0.02 %

Table 1: Part-of-speech distribution: Book 1 *of Plato's Republic and* A day in spring *by C. Kosmač*

Comparative Analysis of English and Slovak translations of Plato's Republic

Alexandra Jarošová
Slovak Academy of Sciences
Ľudovít Štúr Linguistics Institute
Bratislava, Slovakia
e-mail: sasaj@savba.savba.sk

Text alignment has at least two dimensions:

1) aligned corpora are treated as useful text sources of translational equivalents (with regard to an improvement of bilingual dictionaries);

2) the alignment itself can profit from some types of equivalents.

Most of the published alignment methods attempt to identify correspondences in parallel texts at the sentence level. As for word alignment, M. Kay and M. Roescheisen (1993, 121) state that it is relatively easy to establish correspondences between such words as proper names and technical terms. In the context of Plato's Republic we can ask which words in such a non-informative text have the status of "technical terms".

In this contribution I try to verify J. Sinclair's hypothesis about the key-concepts of a particular text as being "translation-protected expressions" - TPEs (TELRI WG 9 documents, 1996). Besides that I will try to expand the notion of TPE with respect to three groups of translation equivalents (TE) proposed by W. Teubert:

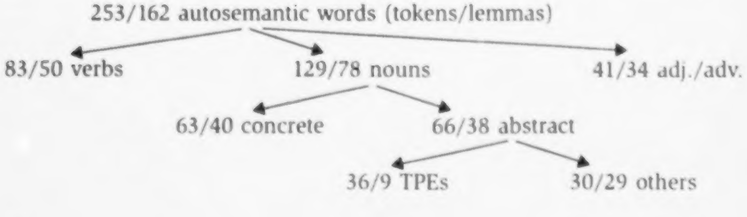
A. TEs found in a bilingual dictionary (BD);
B. TEs not found in BD and not regarded as appropriate;
C. TEs not found in BD but regarded as suitable.

The C list contains, as a rule, context-sensitive but recurrent TEs. W. Teubert proposes to carry out the work on these TEs by formulating rules for their generalization. The objective of W. Teubert's proposal is "to computerise the detection of those TEs which have not yet been recorded in bilingual dictionaries" (TELRI WG 9 documents, 1996). We believe that one-to-one correspondence between a specific word and its equivalent (represented predominantly in the A list) is not a basic issue of the interlingual equivalence problem. The main equivalence type is a partial equivalence which can be resolved on the word combination level. The C list contains, as a rule, a number of different multiword units associated with a source language word, or the synonyms of prototypical equivalents corresponding to source language collocations.

The result of our analysis is influenced by the fact that the two compared texts are translations of a third one and that the English translation is about 100 years older than the Slovak one. On the other hand the analysis of uneasily alignable texts can serve as a basis of detection of many different problems.

358 words (tokens) from the analysed piece of English text (2010 words) have their prototypical (i.e. found in dictionary) equivalents in the Slovak version (Dujnic,1996). Besides that there are 50 instances of correlation between proper names. Thus 407 tokens could be aligned (if lemmatized) by the existing bilingual dictionary and by an ad hoc created dictionary of proper names (e.g. Cephalus - Kefalos, Glaucon - Glaukon, Themistocles - Temistokles). This means that 20.3% of tokens in these parallel texts are alignable (approximately every sixth token). 104 instances of 357 prototypical equivalents (tokens) are grammatical words and 253 instances are autosemantic words. In spite of the high frequency of grammatical words in any text they do not represent an important group in the prototypical equivalence domain. This is caused partially by their high polyfunctionality and their position in grammatical systems of the respective languages, e.g., articles, personal and possessive pronouns, "there + to be" pattern in an English text have very low or no correspondence in a Slovak one and conversely, the Slovak reflexive particle "sa" and auxiliaries in the simple past tense have no counterparts in an English sentence. Nevertheless there are some grammatical words which demonstrate a relatively high degree of mutual bilingual correspondence. In the compared texts such words are the conjunctions: Eng. "but" (24) - Slov. "ale" (19), Eng. "if" (9) - Slov. "ak" (7), Eng. "or" (10) - Slov. "alebo" (12) and the preposition Eng. "with" (14) - Slov. "s" (12). Grammatical words are often components of fixed word combinations and these combinations (triples, quadruples etc.) can be an object of alignment (c.f. Karin Cvetko-Rasmussen's presentation at the WG9 meeting in 1996, Mannheim).

The analysis of the 253 autosemantic tokens (162 lemmas) gives us the following picture. About one half of the number are nouns (129 tokens / 78 lemmas), 83 tokens (50 lemmas) are verbs and 41 tokens (34 lemmas) are adjectives and adverbs. 63 instances (40 lemmas) of the 129 nouns have a more or less concrete meaning and 66 instances (38 lemmas) have an abstract one. 36 abstract nouns (9 lemmas) could be regarded as equivalents of translation protected expressions.



Within the scope of the analysed fragment (Socrates' dialogue with Cephalus about old age) the key-concepts have been determined partially by their frequency and partially by their importance for the given topic. Some of the key concepts have not been translated by Slovak prototypical equivalents in all their occurrences but in the majority of them. From three tokens of the lemmas "life", "justice", "true" and "truth" in the English text only two tokens have prototypical equivalents in the proper Slovak sentences. (With this frequency they are included in the A list). In the group of abstract nouns regarded as key-concepts we find the following distribtion:

age (15) – staroba (14), vek (1)
wealth (4) – majetok (2), bohatstvo (2)
hope (3) – nádej (3)
pleasure (3) – pÔžitok (1), rozkoš (1), radosť(1)
justice (2) – spravodlivosť (2)
true (2) – pravdivý (1), pravda (1)
truth (3) – pravda (2), pravdivosť (1)
life (2) – život (2)
youth (2) – mladosť (2)

The key words "age", "justice", "hope", "youth", "life" are true translation protected expressions. The key words "true", "truth" and "wealth" each have 2 equivalents (the problematic case "pleasure" has 3), but these equivalents are within the scope of prototypicality. So these words could be regarded as translation protected to some degree (as a rule within 2 equivalents).

The frequent words "man" (9 occurrences) and "men" (7 occurrences) do not behave as completely translation protected expressions, but there is a specific tendency in the distribution of equivalents.

We have included (with a question mark) "man/men" in the list of concrete nouns. These nouns represent a group of words that are close to TPEs:

1 arms
1 body
2 brother
1 chair
1 child
1 children
1 city
1 cloak
1 court
1 evening
3 father
2 festival
2 friend
1 garland
1 god
2 goddess
2 grandfather
1 head
1 heir
1 home
1 horse
3 man
5 men
3 master
2 money
1 night
1 poem
2 poet
1 procession
1 proverb
1 race
2 relations
1 sacrifice
5 son
1 supper
1 threshold
1 torch
1 way
1 word
1 world

A comparison of the first three sentences in three different English translations of Plato's Republic (Jowett 1901; Waterfield 1994; Bloom 1991) has shown that these translations have very few identical words, but almost all identical words occur in places where the prototypical equivalents occur in corresponding Slovak sentences ("went", "down", "yesterday", "son", "goddess", "festival", "procession", "run", "wait", "him", "us"). Nouns are represented here by the type with concrete meaning. Another relevant observation: 11 concrete nouns occurring in Jowett's translation do not have the prototypical equivalent in the Slovak version (e.g., "inhabitant", "servant", "house", "horseman", "chair", "authors", "parents"). In Bloom's translation nine words from that group are replaced by other concrete nouns (e.g. "native", "slave", "home", "stool", "poets", "fathers"), which are in prototypical equivalence relation with the corresponding Slovak words.

So the percentage of concrete nouns in the A list can be higher depending upon the degree of similarity between the two translations.

The most frequent group of verbs in the A list are, as could be expected, reporting verbs (17 say, 4 ask, answer, tell). Various forms of the lemma "to say" occur in the relevant English fragment of Republic 29 times, out of which seventeen instances were "translated" by the prototypical Slovak equivalent ("povedať"). Individual grammatical forms of this verb are not equal from the prototypical equivalence point of view. Slovak equivalents of the other reporting verbs (e.g. "to reply" and "to tell") are candidates for inclusion in the C list.

A more detailed characteristics of the C list will be the subject of my contribution to the Kaunas Seminar.

CONCLUSIONS

The A list (English words with prototypical Slovak equivalents) contains 162 autosemantic lemmas (more precisely pairs of lemmas). Almost half of them are nouns (78). Two groups of nouns (9 + 40) behave as translation protected expressions (analogues to "technical terms" in informative texts): **1.** abstract nouns expressing the key-concepts of the given text (9 words) and **2.** nouns naming concrete objects ("court", "garland", "poem") or persons ("servant", "brother", "poet') of the real or an imaginary world ("goddess"). In the group of verbs, reporting verbs play a similar role in this particular text.

DICTIONARIES AND TEXT SOURCES

Dujnic, M. 1996. Moderny anglicko-slovensky slovnik. Gardenia Publishers. Bratislava
Plato. "Republic". 1994. Translated by R. Waterfield. Oxford University Press. Oxford - New York.
Plato. "Stat". 1990. In "Dialogy II." Translated by J. Spanar. Tatran Publishers. Bratislava.
"The Republic of Plato." 1901. Translated by B. Jowett. The Colonial Press. New York.
"The Republic of Plato". 1991. 2nd ed. Translated by A. Bloom. Harper Collins Publishers.

LITERATURE

Cvetko-Rasmussen, K. 1996. TELRI internal reports and documents.

Kay, M. - Roescheisen, M. 1993. "Text-Translation Alignment." Computational Linguistics, 19 (1), 121-142.

Sinclair, J. 1996. TELRI internal reports and documents.

Teubert, W. 1966. TELRI internal reports and documents.

Plato's Republic: Some comments on translation

Ann Lawson
Corpus Research, University of Birmingham
Birmingham, Great Britain
e-mail: a.e.lawson@bham.ac.uk

All the texts of Plato's "Republic" used in the "Joint Research" project have been translated from a third language, the original Ancient Greek, thus eliminating potential mistranslations, poor or clumsy translations, omissions, explanatory additions or archaisms found in one of the versions and carried over to another. We can only talk, then, in this case of the Source Language as the Greek and Target Language as any of the myriad versions that the project has collected. As Wolfgang Teubert points out ("Comparable or Parallel Corpora?" in IJL vol 9, no 3, Sept 1996, p 239), the two Target Language versions, here English and German, each have more in common with the original source version than with each other. As I am unfamiliar with Ancient Greek, I have to treat the modern versions independently[1]. This has both disadvantages and advantages, as it makes analysis harder yet provides a more authentic framework. After all, readers of works in translation generally read translations because they lack a sufficient command of the language to read the original.

As Tognini-Bonelli points out ("Towards Translation Equivalence from a Corpus Linguistics Perspective" in IJL vol 9, no 3, Sept 1996), translation presupposes what we may call 'displaced situationality'. That is, the linguistic context (or 'co-text') is displaced because it differs in the two languages, and the situational features will differ as a different culture, situation and participants is referred to. This is still more an issue with this text because it deals with an era long past and inevitably much of the vocabulary is archaic and unfamiliar to the modern reader. The text and its language are also abstract and philosophical by nature, making the translation process yet harder as the precise meaning of concepts is more difficult to pin down and explain neatly. It is clearly also an issue whether the modern translator chooses to make the text more accessible to the reader by exchanging unfamiliar words or phrases for more modern approximations (see below).

---

[1] The Jowett Translation used is the third edition of Jowett's multi-volume Plato, produced in 1892 and subsequently used as the basis for most critical editions in English. The more recent English translation was made in 1969 by Paul Shorey, Cambridge, MA, Harvard (University Press; London, William Heinemann Ltd.). Both are available for use by TELRI partners on the Mannheim ftp-server, as is the German text.

I will use the analysis of one word and its accepted translation equivalents to investigate some aspects of translation highlighted by these parallel texts.

The word "knowledge" occurs 152 times in Jowett's translation, compared with just 95 times in the more recent Shorey translation. In the German, a search for the generally accepted translation equivalents "Wissen", "Kenntnis" and their plural and genitive forms found only 67 instances in total. Learners of German are generally taught that "Wissen" is used in the context of "knowledge about something" or "knowledge of how to do something", while "Kenntnis" is used for acquaintance with people and facts. Given these options provided by traditional paper bilingual dictionaries, the discrepancy in the frequencies seems surprising and curious.

A full statistical analysis was not possible at this stage due to restrictions in the software I was using. However, even a preliminary examination of the instances of "knowledge" which were translated differently uncovers some interesting findings.

[340e][2]

Shorey: For it when his **knowledge** abandons him that he who goes wrong goes wrong – when he is not a craftsman.

Jowett: they none of them err unless their **skill** fails them, and then they cease to be skilled artists

German: Wo sein **Fachwissen** auslässt, dort irrt der Irrende, worin er also nicht Fachmann ist

[350a]

Shorey: Consider then with regard to all forms of **knowledge** and ignorance and ignorance whether you think that anyone who knows would choose to do or say...

Jowett: And what about **knowledge** and ignorance in general; see whether you think that any man who has **knowledge** ever would wish to have the choice...

---

[2] This number refers to the folio markings commonly found in the text, and present in both the English and German versions.

German: Was **Fachwissen** und Dilettantismus allgemein anlangt: Will ein Fachkundiger vor einem andern Fachkundigen etwas voraushaben...

It would appear, then, that "Fachwissen", the specialized knowledge of a subject, can be considered a suitable translation equivalent of "knowledge" in certain contexts. Since it occurs only 5 times throughout, it cannot, however, explain the numerical discrepancy.

[366d]
Shorey: (he) is aware that a man..., having won to **knowledge**, refrains from it.
Jowett: ...or has attained **knowledge** of the truth...
German: ...oder sich aus tiefer **Erkenntnis**} seiner ablehnt...

"Erkenntnis" would normally be translated as "recognition" or "realisation", and thus provides another translation equivalent not generally offered by conventional means.

[409c]
Shorey: ...by the instrument of mere **knowledge** and not by the experience of his own
Jowett: ...**knowledge** should be his guide, not personal experience
German: ...kraft seines **Geistes**, nicht der persönlichen Erfahrung

[411e]
Shorey: ...it seems that there are two arts which I would say some god gave to mankind, music and gymnastics for the service of the high-spirited principle and the love of **knowledge** in them
Jowett: And as there are two principles of human nature, one the spirited and the other **the philosophical**
German: Für diese beiden Anlagen gab, so glaube ich, ein Gott dem Menschen die zwei Künste der Musik und der Gymnastik, fűr das Mutvolle und **das Geistige** in ihm

"Geist" is a highly problematic word and indeed concept for English, for which there is no straightforward translation equivalent. The term can refer to the mind, the spirit or intellect, as well as a ghost, its cognate in English. It seems highly probable that as there is no such encompassing yet abstract term in English, this is a translation solution that may well not occur to a translator instinctively. The benefit of using translations originating from Greek is that the natural phraseology and word-choice is less affected by the influence of the English or German, which are rather similar languages in many ways. Any translator from German into English would have great difficulty with the adequate explanation of "Geist".

These are but some of the examples presented in more detail in Ljubljana. They indicate, however, how a network of TEs can be built up using parallel texts, which includes many which would never normally connect in traditional dictionaries. These options could be presented to the translator in context, with examples of their previous translations.

The translator must decide whether to translate "authentically", as it were, using terms similar to the ones Plato used, or whether to translate those terms into similar equivalents which the modern reader would understand. One such example is the translation of the Greek word for a voluminous cloak. In German, this is rendered as

German: Der Diener fasste mich am **Mantel**
Jowett: The servant took hold of me by the **cloak** behind
Shorey: The boy caught hold of my **himation** from behind

Really this last version could better be termed a non-translation, since the transliteration of the ancient Greek word is the same. It clearly referred originally to a garment unlike that familiar to a modern reader, but the translator must decide whether an approximation of meaning is more desirable than a potential misunderstanding or non-understanding. As it stands, "himation" presents an almost insuperable barrier to most modern readers, as it is unknown to all but scholars of Ancient Greece (who are unlikely to be reading in translation!) and it is found in few dictionaries.

The same aspect of choice and arbitrariness holds true for the translation of idiom. [329a]sees the quoting by a character of an old proverb, undoubtedly different in the original, but rendered as follows in the versions examined here:

German: Oft kommen wir Gleichaltrigen zusammen und bestätigen das alte Sprichwort[3] .
Jowett: Men of my age flock together; we are birds of a feather, as the old proverb says
Shorey: For it often happens that some of us elders of about the same age come together and verify the old saw of like to like.

Here, it would appear that the Jowett version gives the modern reader the best concept of what the original saying expresses. There is in fact no such

---

[3] "Gleich und gleich gesellt sich gern" might be better alternative translation.

"saw" (another highly unfamiliar word for the modern English native speaker!) as "elders of the same age coming together". A much more natural and understandable proverb is the well-known "Birds of a feather flock together", so well known that it is often referred to only partially, as in this example, in the understanding that the reader will be able to recreate the complete proverb.

The analysis of parallel corpora underlines the creative nature of translation, as the translator adapts the text according to their understanding, ideology and their target audience. All these factors mean that any attempted translation between the German and English texts would inevitably result in a very different version of the text than those produced by the individual translations from the Ancient Greek.

For a translator, it would be useful to be able to select style, genre and era of a text. Of course, the majority of translations involve technical, legal or administrative language, and literary translations will almost certianly always be best undertaken with at the very least input from experts in the field, perhaps in conjunction with a critical edition. Even so, some translation aids to offer suggestions and examples would be useful to even the most experienced literary translator. For more technical translations, it is highly likely that options of choosing subject-area, style etc would be desirable. Much terminology alters meaning at least subtly and sometimes violently from subject-area to subject-area, and even a careful translator can easily be caught out. Further work on parallel texts can clearly assist in the development of such tools and methodologies.

Structure Markup in CES and Preliminary Statistics on Romanian Translation of Plato's "Republica"

Dan Tufiş, Ştefan Bruda
Romanian Academy of Sciences
Bucharest, Romania
e-mail: {tufis, bruda}@valhalla.racai.ro

The Romanian version of Plato's "Republic" was translated by Andrei Cornea, taking as reference the Brunet edition (1968) published by Oxford Press. This first integral translation of Plato's work was supervised by Constantin Noica (who also wrote the Preface) and Petru Creţia two of the most reputed Romanian scholars, thus warranting an accurate translation and interpretation of

"Republic". The book has been keyboarded and typos have been carefully removed by spell-checking and lexical lookup in a large electronic lexicon.

In SGML marking-up, we used the TEI conformant CES-1 (paragraph level) encoding schema (see details on CES encoding at *http://www.cs.vassar.edu/CES/*), with a few sub-paragraph elements included. With a word count of 131064, the tag usage of our encoding is the following: body=1, div=32, head=32, hi=339, p=4301, q=4233, quote=39, name=700, poem=6, l=11, xptr=1341 and corr=265. The correction mark-up (corr) was used to structurally identify pieces of text which were not physically in the original but could be inferred with maximum probability. These translator's additions, marked in the printed version as well, are in most cases modifiers (adjectives, relative clauses, etc.) which if removed does not affect the grammaticality of the remaining text. The marginal notes were encoded as xptr elements with unique identifiers (the uniqueness has been achieved by appropriate concatenation of the prefix R-for Republica and the numeric and literal marginal notes: R327a, R327b,... R621.d). Comparing our CES encoding with the TEI-lite encoding of the English version (Benjamin Jowett, P. F. Collier &Son. New York, 1901 edition) a wild discrepancy has been noticed in terms of paragraph (p) and quotation (quote and q) marking: 4298 paragraphs in Romanian versus 40 in English, 4272 quotations (4233 q and 39 quote) in Romanian versus 162 (only q) in English. This was due to the very different rendering of the printed books in the two languages. By means of the on-line translator Fred (service available at the address *http://www.oclc.org/fred/docs/translations/translate.html*) the SGML encoding of Republica was converted into HTML (a sample can be seen at *http://nl.ijs.si/telri-wg5/Republic/plato-ro.sample.html*).

The text of the electronic version of "Republic" was segmented by means of a tokenizer, part of the tool-set implemented within the MULTEXT project (see the MULTEXT home-page at *http://www.lpl-univ-aix.fr/projects/MULTEXT/*) with the language resources we developed within the MULTEXT-EAST project (see the MULTEXT-EAST home-page at *http://nl.ijs.si/ME*). The segmenter is a language independent and configurable processor used to tokenize an input text, given in one of the three possible formats: plain text (without any mark-up), a normalized SGML form (nSGML) as output by another MULTEXT tool (MTSgmlQl) and a tabular format (also specific to MULTEXT processing chain). The output of the segmenter is a tokenized form of the input text, with paragraph and sentence boundary marked-up. Punctuation, lexical items, numbers and several alpha-numeric sequences(such as dates and hours) are annotated with various tags out of a hierarchy class structured tagset. A lexical item may be an orthographic word (delimited by spaces and/or punctuation), a part of an orthographic word (clitics are split up), an abbreviation or a compound made up of two or more orthographic words. The language specific behavior of the segmenter is driven by several language resources (abbreviations, compounds, clitics, etc.). The general behavior of the segmenter (valid over several languages) can also be parametrized by means of external resources such as definition for space and punctuation, number orthography, sentence delimiters, etc.

Once the input text is tokenized, a dictionary look-up procedure, can be invoked to assign each lexical token all its possible morpho-lexical interpretations (see Figure 1, first column). This procedure was incorporated into a special XEMACS mode (mtems-mode, due to Tomaz Erjavec) in order to take advantage of the editing facilities of XEMACS. By knowing the significance of the morpho-syntactic codes we used, the mtems-mode allows a user to manually disambiguate the segmenter's output (see Figure 1, second column).

spre	Spsa #	Spsa #
a	I Qn Spsa Tsfs Va--3s #	Qn #
mă	I Pp1-sa--------w #	Pp1-sa--------w #
ruga	Vmnp Ncfsry Vmii3s Vmm-2s #	Vmnp #
zeiţei	Ncfsoy #	Ncfsoy #

Figure 1: Tokenized text with ambiguous annotation (first column) and final annotation (second column)

The human disambiguator is shown (at request) the significance of the MSD codes, having several editing possibilities. These codes (MSDs) are conformant with the MULTEXT-EAST linear encoding specifications (see the Morphosyntactic Encoding Description at *http://nl.ijs.si/ME/Docs/Multext/multext-lexical-encoding*), which themselves, represent an extension of the EAGLES proposal for morpho-syntactic annotation (see ; while EAGLES considered only Western languages, MULTEXT-EAST specifications cover two more language families: Slavic and Ugro-Finic. Romanian, being a Romance language was in principle covered by EAGLES specifications. Manual disambiguation was, as one can imagine the most time consuming phase in our processing of "Republica". It was done by a professional linguist and therefore, presumably, it is a clean and valuable language resource (among others we plan to use this disambiguated text, together with other fictional texts, for building language models and performance evaluation of tagging with several tagsets, input data for grammar induction etc.).

With respect to automatic tagging, the MSDs set is mapped, during the process of building the language model, onto a corpus tagset by means of an external resource, so by modifying only this file, we could experiment with several tagsets without modifying the disambiguated texts. A semi-automatic procedure, allows for generating from the MSD several tagsets, according to criteria specified by the designer. The Romanian MSD set has 550 codes which is definitely too much for a satisfactory performance of a HMM tagger. On the other hand, we would like to keep the corpus tags as informative as possible. This is why the first try will be carried out with a tagset containing 187 codes.

Based on the manually disambiguated texts (about. 200.000 lemmatized and MSD annotated words) we extracted some counts and frequencies. Although these statistics are not supported by enough data (we plan to extend the analysis on several million of words automatically lemmatized and disambiguated, as soon as the language models we mentioned above would ensure a reasonable precision). Here, we will only refer to the data extracted from a chunk (about 20.000 words) of text extracted from Plato's "Republica".

Figure 2 lists the most frequent 13 MSDs. This frequency list, as all the others computed based only on the selected chunk of text from "Republica", does not conform with the one based on the whole corpus, but as far as the functional words are concerned, their usage is quite accurately reflected by the partial figures discussed here. The toplist is, not surprisingly, the simple preposition subcategorizing for an accusative NP.

Spsa	0.080944
Rgp	0.0753307
Ccssp	0.0561357
Ncfsrn	0.0326553
Csssp	0.0326553
Qs	0.0296372
Rp	0.0265588
Ncms-n	0.0265588
Ncfsry	0.0257741
Qz	0.0237219
Vmip3s	0.0228164
Afpms-n	0.0228164
Vmnp	0.0226957

Figure 2: Relative distribution of MSDs (first 13 most frequent)

Verb	0.226414
Noun	0.164785
Pronoun	0.129052
Adverb	0.120963
Conjunction	0.0955514
Preposition	0.0867387
Particle	0.0613267
Adjective	0.0514275
Article	0.0351301
Determiners	0.0226354
Numeral	0.0035613
Interjection	0.00229372
Abbreviation	0.00012072

Figure 3: Relative distribution of POS

Figure 3 ranks the distribution of the wordforms according to their part-of-speech. This time, the first position is occupied by the verb (main, auxiliary, copulative), followed by nouns, pronouns, etc. We should mention that what comes under particle count is greatly due to the preposition conjunction "s,", negative adverb "nu" and preposition "a". Therefore, under a traditional classification, the adverb would come in the third position, pronoun in the fourth and so on.

să	Qs	485
şi	Ccssp	432
nu	Qz	393
de	Spsa	347
că	Csssp	319
mai	Rp	262
fi	Vcip3s	224
în	Spsa	222
pe	Spsa	221
avea	Va--3	156
dar	Ccssp	153
se	Px3--a--------w	152
şi	Rp	148
ce	Pw3--r	144
avea	Va--1	140
cu	Spsa	137
care	Pw3--r	131

Figure 4: Distribution of MSDs per lemma

581	Ccssp Px3--d--------w Rp
295	Vcip3s Vmip3s
222	Vmip1s Vmsp1s
194	I Qn Spsa Tsfs Va--3s
191	Csssp Rgp
189	Vmnp Vmip3s Vmm-2s
178	Vmip2s Vmsp2s
154	Vmip3s Pw3--r Dw3--r
154	Afp Rgp
152	Ncms-n Vmip1s Va--3
151	Va--1 Vmip1s Vmsp1s
134	Pw3--r Dw3--r--e
126	Qf Pp3fsa--------w Mcfsrl Tifsr Va--3s
117	Afpms-n Vmp--sm
113	Ncfp-n Ncfson
111	Vmnp Vmii3s Vmm-2s
101	Afpms-n Ncms-n Rgp Spsa

Figure 5: Distribution of ambiguity classes

A few counts of the most frequent lemmatized worforms are shown in Figure 4. One should notice the high position of the verbs "a fi" (to be) in its copulative use and "avea" (to have) in its form.

Other interesting data (Figure 5) were obtained by counting the ambiguity classes in a segmented (but not disambiguated text). Due to space limitation we cannot develop this issue, but these frequencies provided valuable hints concerning mapping MSDs onto corpus tags.

News from TELRI Working Groups

▮ *WG 9 ORGANISING JOINT REASEARCH*

Co-ordinator: John Sinclair

The Joint Research Group of TELRI has the tricky brief of learning how to work together on research, with funding only for the most basic needs. It was therefore necessary to devise projects that did not carry heavy overheads nor make large demands on the time of expert researchers - but which nevertheless were serious academic explorations.

In addition it was seen as an excellent opportunity to use the rich range of languages available among TELRI partners to stress the importance of multilingualism. Now that steps have been taken (eg in the PAROLE project) to provide, language by language, substantial generic resources, the issue arises of relating these to each other. Hence it was decided to build a multilingual "corpus" that empasised breadth of languages rather than overall size; to begin with, a single text. With this resource we could first of all try out how difficult it was to find equivalent texts in the languages from Estonian to Albanian, tackle the problems of putting all the texts into electronic form, and then survey the field to see what tools were available to exploit this resource.

Difficulties and problems there were, but by the halfway mark of the project most had been resolved. We went on to a survey of tools for alignment etc, which gave results that were not very promising. Given the diversity of the languages, the varied types of translation and the range of conventions used for electronic conversion, there was little we could find on the market that was likely to give good results. Although we had just a single text, we tried to imagine it as much too long to be preprocessed or marked up by hand, and concentrated on techniques that were independent of both language and size.

Most alignment software relies on consistency of paragraphs and even sentences across the translation boundary. In our study, this seemed to be very optimistic, and so, in parallel with exploring conventional alignment techniques we investigated some linguistic hypotheses which did not rely on physical alignment. The first reports were presented in Ljubljana.

This is clearly Work in Progress; scholars in very diverse circumstances focusing on the same general problem, and bringing to it their own expertise, research traditions and the individuality of the languages they are working with. Later this year, at the final TELRI seminar, we hope to present a further stage in our research, showing that we have fulfilled one of the principal aims of TELRI - to build a language resources infrastructure within which researchers throughout Europe can work easily.

THE BRIDGE PROJECT

The second project of the Joint Research Group stresses another important aspect of TELRI - working with commercial partners, and participating in product development where success is judged in the market place. For this, partners took advantage of an opportunity to build a family of translated dictionaries. A small modern dictionary - the Cobuild Students Dictionary - was made available by courtesy of the publishers, HarperCollins. This is a monolingual English Dictionary whose definitions are written in complete sentences, thus simplifying the translation process. A translation is already published, in Brazilian Portuguese, and others are in progress; members of TELRI were asked if they would like to arrange for a translation to be made and published in their own languages. This aroused considerable interest and several ventures have started. At each TELRI event another step is taken in co-ordinating the work and passing on the experience gained; in Ljubljana the English and Portuguese texts were compared.

As each new translation is added, each Bridge partner gets a bonus, because another set of multilingual dictionaries is now possible. So there is every reason to get beyond the first hurdle, of each partner doing a translation. Also there is good reason to believe that a machine-usable lexicon of considerable generality could be derived from this pool of translations, leading to another stage of exploitation, one that could hardly be contemplated before the pool was in sight.

There are all sorts of problems to be overcome in this project; as well as the considerable linguistic ones of establishing conventions in a new kind of lexicography, there are serious economic and commercial issues to be resolved; each partner meets a unique set of circumstances and tries to resolve the difficulties. The Bridge project looks set to go on for some time, and owes a great debt to TELRI for support in inaugurating it. The presence at the Ljubljana meeting of the Slovenian publisher who will work with TELRI on the Slovenian translation was evidence that the objective of involving commercial interests is gradually being met.

Prof. John M. Sinclair, *Tuscan Word Centre, Azienda Casanova, 409 Vellano, 51010 Pescia (PT), Italia, Phone:+39 572 40 92 51; fax 40 92 53*

On the Lexicons

The Historical Dictionary of Hungarian

Julia Pajzs
Research Institute for Linguistics,
Hungarian Academy of Sciences,
Budapest, Hungary
e-mail: pajzs@nytud.hu

The project for compiling the Historical Dictionary of Hungarian originally started more than onehundred years ago. For several decades enthusiastic men of letters have been collecting the traditional dictionary slips. This collection now contains about 5 million slips, but unfortunately they are not even alphabetised correctly, so it is very difficult to use them. Therefore in 1985 the Academy of Sciences decided to start a new project in which the first task was to create a computerised corpus of the Hungarian language. At that time there was hardly any experience in collecting and using of corpora, therefore the original ideas on the size of the planned corpus and the way of collecting it seem rather old fashioned nowadays. It was hoped that a corpus of 10 million running words from 5 centuries could properly cover the vocabulary of the language. For this purpose literary historians selected the sample texts from each century, which are being keyboarded manually. Since the sample texts are usually very small excerpts (just a few pages continuously from here and there) neither optical scanning nor the use of typesetting tapes proved to be efficient enough. Keyboarding of the texts from the earlier centuries meant additional difficulties even for the human typists, so we rather concentrated on the texts from the 19-20 centuries. Although the size of the current on-line corpus is nearly twice as large as it was originally planned (17 million words on-line, another 3 million only keyboarded but not yet controlled, corrected), we had to cope with the fact that it is still hopelessly too small for the compilation of a historical dictionary. The currently available material could more or less satisfy the needs of a new corpus based dictionary, but not the planned, OED like one. To give an idea on the usability of the corpus here are the main results of the lemmatised frequency list: the number of the different lexemes was 165.442 (not taking into consideration the homographs), only 64.090 occurred at least 3 times. While we are trying to compile the draft entries we often do not find enough examples, or they do not illustrate the different possible meanings properly, not to mention the change of the meaning throughout the centuries.

Thus after 12 years we still feel that we are at the very beginning of our historical dictionary project. During this time we have developed and purchased some software tools for retrieving our corpus and analysing it. The ready made Open Text (earlier PAT) software was bought some years ago, we have made a Hungarian interface for it with which anybody can retrieve the occurrences of the running words or expressions (via telnet). The search can be limited according to the date of writing, literary forms or authors. The results can be saved and e-mailed. We will further develop this interface in two directions: one for the external users, and one for the lexicographers and other linguists in the institute.

A morphological analyser program was also written in collaboration with MorphoLogic Ltd. (presented on the first TELRI seminar). The program segments the running words into lexemes and suffixes (prefixes). After the analysis we can retrieve the lexemes with the Open Text software. The program cannot disambiguate the homographs, because it is a pure morphological analyser. Therefore we are carrying out research in this field in the framework of GRAMLEX COPERNICUS project. The research is three-directional, we compare different methods: a pure statistical tagger (HMM), a statistical-grammatical tagger which uses local rules for the decisions, and we started to write a syntactic analyser for the same purpose. The result of these will be compared and evaluated. For writing the dictionary entries we use the WriterStation program which was dedicated for SGML text editing. Once you have decided the DTD of your dictionary it is fairly easy to develop an application in this framework. However, the version we own is relatively old (it was bought for us 4 years ago), and we are not satisfied enough with it to buy a new version. We rather intend to find a good and inexpensive tool for this purpose in the near future.

We continue to enlarge our corpus both by traditional keyboarding and by obtaining electronic texts from publishers. The main obstacle of our project is the usual one: the lack of money and manpower. The staff of the project is rather small considering the task (6 researchers and 2 keyboarders), and we have no reason to hope that it can be sufficiently increased in the near future. However we are trying to do our best in the compilation of a corpus based dictionary of Hungarian.

BIBLIOGRAPHY

Pajzs J.: *Creating a Historical Dictionary of Hungarian with the Aid of Computer* T. Magay - J. Zigany: BUDALEX '88 Proceedings Akademia Kiado Budapest 1990. 559-563.

Pajzs J.: Realisation assisteé par ordinateur de grands dictionnaires français et hongrois Cahiers détudes hongroises 3/91 Centre Interuniversitaire détudes Hongroises Université Paris III. Institut Hongrois de Paris, 47-54.

Pajzs J.: *The Use of a Lemmatized Corpus for Compiling the Dictionary of Hungarian* In Using Corpora Proceedings of the 7th Annual Conference of the OUP & Centre for the New OED and Text Research. University of Waterloo Centre for the New OED, 1991. 129-136.

Pajzs J.: *Project Report on the Historical Dictionary of Hungarian.* in: KIEFER F. - KISS G. - PAJZS J., eds Papers on Computational Lexicography and Text Research. Proceedings of COMPLEX '94 Budapest 1994. 205-214.

Proszeky G. - Tihanyi L.: *A Fast Morphological Analyser for Lemmatizing Corpora of Agglutinative Languages.* Papers in Computational Lexicography. Proceedings of COMPLEX '92. Edited by F. KIEFER, G. KISS J. PAJZS Budapest 1992. p. 275-278.

Proszeky G.: *HUMOR - A Morphological System for Corpus Analysis.* Proceedings of the first TELRI Seminar in Tihany. Ed. Rettig, H. Budapest 1996. p. 149-158.

Lexicon for a linguistic annotation of Dutch text.

John van der Voort van der Kleij, Truus Kruyt
Institute for Dutch Lexicology, INL,
Leiden, The Netherlands
e-mail: {john, kruyt}@rulxho.LeidenUniv.nl

INTRODUCTION

The Institute for Dutch Lexicology INL started automatic linguistic annotation of present-day Dutch texts in 1992. In the framework of the EC-project *NERC* (Calzolari et al. 1996), linguistic software was developed which automatically provided the words (tokens) of an electronic text with headword (lemma) and Part of Speech (POS) (Panhuijsen et al. 1992). The lemmatizer-tagger, called *DutchTale*, consists of three components: (a) a lexicon, (b) a rule component containing morphological rules and restricted context rules (trigrammodel), and (c) a statistical component. By the lexicon, none, one or more lemma's and POS's are assigned to the tokens in the text. By the rule sets and the statistical component, tokens ambiguous with respect to lemma and/or POS are disambiguated, and many tokens not found in the lexicon (including inflected forms, compounds, etc.) still get a lemma and/or POS.

Improved versions of Dutchtale have been applied to three text corpora consultable via Internet, *INL 5 Million Words Corpus 1994* (Van der Voort van der Kleij et al. 1994), a *INL 27 Million Words Newspaper Corpus 1995* (Kruyt et al. 1995), and *INL 38 Million Words Corpus 1996*. Rather than the lexicon component, the improvements concerned the other two components. However, the lexicon now is to be revised, due to a new Dutch spelling system officially prescribed since August 1996. First, the origin and composition of the present lexicon is described. Then, we will outline which procedures are needed in order to get a lexicon covering texts written in both the former and the new spelling.

LEXICON COMPOSITION

In 1990, INL produced the *Herziene woordenlijst Nederlandse taal* (SDU, Den Haag), a corpus-based extension of the official Dutch spelling guide *Woordenlijst van de Nederlandse taal* (Den Haag 1954). This list includes headwords with additional linguistic information such as gender, meaning (mainly for homographs), flected forms and allowed spelling variants. As the wordlist of '54, containing ca. 68,000 entries (headwords), was not up to date, ca. 25,000 new entries have been selected on the basis of frequency and distribution in the INL corpora. Furthermore new inflected forms have been added to the existing entries.

The extended wordlist file was the reusable source for the *DutchTale* lexicon. However, POS information (essential for a tagger) was not available in this source. Therefore POS has been derived automatically from special formal features in the file, such as gender encodings for the nouns, specific types of inflected forms for the verbs and the adjectives. For the remaining headwords, POS was added manually. Subsequently, all headwords and inflected forms have been written individually with their POS and their canonical form into a new file. This file hence contains an entry list consisting of headwords and inflected forms. Identical forms were reduced and their information was cumulated. As a consequence, many entries were provided with several POS and/or several headwords. For some lexicon entries, the original ambiguity was pruned according to lexical preferences, in order to prevent superfluous ambiguities. For example, the type ‚heeft' may either be used as an inflected form of the verb 'hebben' (to have), or as the obsolete puristic noun (lift); the latter option was pruned. As verbs had only two or three inflected forms in the original wordlist (an editorial restriction), a morphological expansion of inflected forms has been carried out by means of a program which generated additional inflected forms. Some irregular verbs needed manual intervention.

Together with a collection of proper names and abbreviations/acronyms, the lexicon contains 230.000 entries.

LEXICON REVISION

In 1995, INL produced the new official Dutch spelling guide *Woordenlijst Nederlandse taal* (SDU, Den Haag), a corpus-based extension of the spelling guide of 1954, so as to implement the Dutch spelling reform of 1995 (see Kruyt & Van Sterkenburg 1995). About 56,000 new entries were added and about 14,000 obsolete entries were removed. The database files for this product provided our second reusable source.

First, the obsolete headwords and their inflected forms were removed from the *DutchTale* lexicon. Next, the headwords in the lexicon which were affected by the spelling reform (marked as such in the database files), had to be respelled. These headwords, together with their inflected forms, had to be added in the *DutchTale* lexicon in the new orthography. Before adding them, a check was performed on the lexicon to verify whether they were already present or not. About 31,000 new headwords and their inflected forms had to be added, both in the new and in the old orthography. For the latter, it was necessary to respell the new inflected forms into the old orthography, as the database files did not contain their old forms. This also applied to the automatically expanded inflected forms of new verbs. Again the addition was preceded by a check on the lexicon to verify the absence of the newly introduced entries.

Geographic names and the words in which they are incorporated, form a separate category. Their orthography was established by a special committee and printed in a separate publication. Incorporating this list into the *DutchTale* lexicon requires considerable efforts, as one has to work through the index of the report and to search for all possible variants of the geographic names.

With the resulting *DutchTale* lexicon we are able to cover the 110,000 headwords of the latest spelling guide and their frequent inflected forms, both in the old and in the new orthography. Moreover, the coverage of our lexicon has been raised by the addition of new current headwords.

Our lemmatizer-tagger *DutchTale* consequently is capable to handle Dutch texts published since 1954, irrespective of whether they are written in the latest or in the older spelling.

REFERENCES

Calzolari, N., M. Baker, J.G. Kruyt (eds.)(1996), *Towards a Network of European Reference Corpora, Report of the NERC Consortium Feasibility Study.* Linguistica Computazionale XI. Giardini Editori e Stampatori, Pisa.

Kruyt, J.G. & P.G.J. van Sterkenburg, A New Dutch Spelling Guide. In: H. Rettig (ed.) *Language Resources for Language Technology, Proceedings of the first European TELRI Seminar*, 133-141.

Kruyt, J.G., S.A. Raaijmakers, P.H.J. van der Kamp & R.J. van Strien, On-line Access to Linguistically Annotated Text Corpora of Dutch via Internet. In: H. Rettig (ed.) *Language Resources for Language Technology, Proceedings of the first European TELRI Seminar*, 173-178.

Panhuijsen, M., J. van der Voort van der Kleij & P. Wagenaar (1992), Automatic Lemmatization Experiment - An explorative study, *INL Working Papers* 92.02.

Van der Voort van der Kleij, J., S. Raaijmakers, M. Panhuijsen, M. Meijering & R. van Sterkenburg (1994), Een automatisch geanalyseerd corpus hedendaags Nederlands in een flexibel retrievalsysteem. In: L.G.M. Noordman & W.A.M. de Vroomen (red.) *Informatiewetenschap 1994, Wetenschappelijk bijdragen aan de derde STINFON-conferentie*. Tilburg, 181-194.

Some interesting events – past and future

A New Impetus For The Study Of Computerized Corpora In Prague

Eva Hajičová
Institute of Formal and Applied Linguistics,
Charles University,
Prague, Czech Republic
e-mail: hajicova@ufal.ms.mff.cuni.cz

The most recent major event in the domain of natural language processing in the Czech Republic is the establishment of the LINGUISTIC DATA LABORATORY of the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague.

The preparation of a complex project called "Czech Language in the Age of Computers", which started in 1996 as a joint research project of seven university departments and institutes (of Charles University in Prague and Masaryk University in Brno; the principal investigator is the Institute of Formal and Applied Linguistics) and the aim of which is to create a solid base for a versatile computerized processing of Czech language serving both as a multifaceted source of material for empirical and theoretical linguistic research as well as a base for manysided applications in the domain of text processing and information retrieval, has brought together a strong group of linguists, computational linguists, lexicologists and computer scientists who have been tied by close working contacts for quite a long time. The complex nature of the project makes it possible to integrate into the research also the students and postgradual students both of philosophical faculties and of the computer science institutes.

However, most of the project partners are faculty members of Czech universities, who have teaching and other heavy duties, and the complex character of the task requires to develop both rule-based (symbolic) and empirical (stochastic) methods of large-scale language data processing and to pay attention both to (written) language and to speech. Thus the necessity has arisen to establish a research laboratory fully concentrated on the introduction of the new aspects into the already well-established research. A support of the Czech Ministry of Education has made it possible to found the LINGUISTIC DATA LABORATORY of the Institute of Formal and Applied Linguistics, to equip the Lab with modern computers and software tools and to put together a small team of young enthusiasts broadening the scope of interests in two respects: first, to focus attention on the combination of symbolic and statistic methods of natural language processing (in order to avoid the disadvantages of them) and on the investigation of the possibilities of the application of methods based on data-driven learning, and, second, to apply these methods not only to the analysis of written text but also to speech analysis. In principle, these strictly mathematically based methods are independent of the concrete language; this opens space for mutual international cooperation, both with regard to the use of the data and the results. However, with view of the specific features of Czech as a language typologically different from the languages to which these methods have been mostly used up to now (rich inflection, word order based on other than grammatical principles), it can be expected that the research carried out in the Lab may bring interesting results for the broader Computational Linguistics and AI communities.

Vilém Mathesius Lectures Series 11

Prague, November 10-21 1997
The Vilém Mathesius Centre for Research and Education in Semiotics and Linguistics

- *Sue Atkions* (Great Britain):
  Frame Based Lexicography
- *Charles J. Fillmore* (USA):
  Frame Semantics and the Lexicon
- *Barbara Grosz* (USA):
  Issues of Discourse Analysis
- *Jacob Mey* (Danmark):
  Pragmatic Acts
- *Paolo Ramat* (Italy):
  Linguistic Categories and Linguists´ Categorizations
- *John R. Ross* (Canada):
  Poetics and the Grammar of Space

- *Arnim von Stechow* (Germany):
  The Formal Semantics of Tense, Aspect, and Mood in Subordinate Constructions in German
- *Mark Steedman* (USA):
  The Syntactic Interface
- *Dean S. Wotrh* (USA):
  Diachronic Interaction of Related Languages: Diglossa, bilingualism, or?

The organizers still expect confirmation from Barbara H. Partee, Janet Pierrehumbert and Helmut Schnelle.

Among the invited Czech lecturers are František Čermák, Miroslav Červenka, Eva Hajičová, Oldřich Leška, Jarmila Panevová, Jaroslav Peregrin, Vladimír Petkevič and Petr Sgall.

PARTICIPATION:

The participation fee for VMC 11 is USD 350, which includes tuition fee, accommodation, and lunches. A limited number of grants is available for students from Central and Eastern Europe. These grants cover the tuition fee, accommodation, and lunches (not travel expenses).

FURTHER INFORMATION:

Mrs. Libuše Brdičková
Institute of Formal and Applied Linguistics
Malostranské nám. 25
118 00 Prague 1
Czech Republic

phone ++420-2-2191- 4278
fax: ++420-2-2191-4309
e-mail: {brdickov, hajicova}@ufal.ms.mff.cuni.cz

IMPORTANT DATES:

Application for grants:	**31st of May, 1997**
Notification(grants):	**31st of July, 1997**
Paid registration:	**31st of October, 1997**

List of Participants*:

ANDERSEN Poul	*e-mail:* m764@eurokom.ie
BECI Bahri	*e-mail:* beci@igjl.tirana.al
BENKO Vladimír	*e-mail:* jazybenk@savba.savba.sk
NEW!!! *tel.*: +421+7 32 36 55, *fax*: +421+7 25 49 56	
BIEN Janusz S.	*e-mail:* jsbien@plearn.edu.pl
ČERMÁK František	*e-mail:* frantisek.cermak@ff.cuni.cz
ERJAVEC Tomaž	*e-mail:* et@cogsci.ed.ac.uk
FISIAK Jacek	*e-mail:* fisiak.plpuam11.bitnet
GELLERSTAM Martin	*e-mail:* gellerstam@svenska.gu.se
HAJIČOVÁ Eva,	*e-mail:* hajicova@ufal.mff.cuni.cz
HLADKÁ Barbora	*e-mail:* hladka@ufal.mff.cuni.cz
NEW!!! *tel.*: +420 2 21 91 42 57, *fax*: +420 2 21 91 43 09	
JAKOPIN Primoz	*e-mail:* primoz.jakopin@uni-lj.si
JAROŠOVÁ Alexandra	*e-mail:* sasaj@juls.savba.sk
NEW!!! *tel.*: +421+ 7 33 17 613, *fax*: +421+ 7 33 17 56	
KRUYT Truus	*e-mail:* kruyt@rulxho.leidenuniv.nl
MARCINKEVIČIENĖ Rúta	*e-mail:* ruta.marcinkeviciene@vdu.lt
OIM Haldur	*e-mail:* hoim@psych.ut.ee
PAJZS Júlia	*e-mail:* pajzs@nytud.hu
PASKALEVA Elena	*e-mail:* hellen@bgearn.bitnet
PENCHEV Iordan	*e-mail:* jpen@bgearn.bitnet
LAURENT Romary	*e-mail:* Laurent.Romary@loria.fr
LAWSON Ann	*e-mail:* a.e.lawson@bham.ac.uk
SINCLAIR John M.	*e-mail:* j.sinclair@bham.ac.uk
SPEKTORS Andrejs	*e-mail:* aspekt@mii.lu.lv
TEUBERT Wolfgang, VOLZ Norbert	*e-mail:* telri@ids-mannheim.de
TUFIS Dan	*e-mail:* tufis@roearn.ici.ro
ZAMPOLLI Antonio	*e-mail:* paula@icnucevm.cnuce.cnr.it

---

*You can see detailed addresses in Newsletter No. 2.

WHAT IS TELRI

The Concerted Action TRANS-EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE (TELRI), is a COPERNICUS project funded by the European Commission. TELRI has a duration of three years (1995-1997). It brings together 22 institutions of 17 European countries (Albania, Germany, Great Britain, Slovakia, Italy, Bulgaria, the Czech Republic, Sweden, Slovenia, Romania, Estonia, France, the Netherlands, Latvia, Lithuania, Poland and Hungary).

TELRI is setting up a permanent network of leading national language and language technology centres in the whole of Europe. It pools existing language resources, corpora, machine-readable dictionaries and lexicons, lexical databases, and generic software tools for the creation, re-use, maintanance, validation , and exploitation of linguistic data. It complements these repositories with newly created multilingual resources, offering a wide range of language data to the NLP community. TELRI is establishing a platform where research and industry meet, exchange resources and engage in product-oriented cooperation.

Links have been established with language centres elsewhere in Europe, with relevant European organizations and ventures, and with local language institutions in other parts of the world.

FOR INFORMATION.

Inquiries about TELRI may be addressed to: **Dr. Wolfgang Teubert,** Institut für deutsche Sprache, P. O. Box: 101621, 68016 Mannheim, Germany, Phone: +49 621 1581 437, Fax: +49 621 1581 415, e-mail: telri@ids-mannheim.de

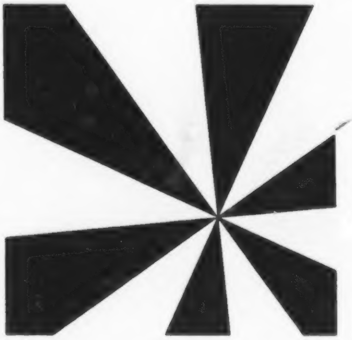
TELRI's WWW Document.

Detailed information about TELRI and its activities is available through the World Wide Web (WWW) at the following URL: **http://www.ids-mannheim.de/telri/telri.html**
Webmaster: Alena Bohmova, e-mail: webadm@smetana.ms.mff.cuni.cz
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Malostranské nám. 25, 118 00 Prague 1, Czech Republic

TELRI

TRANS EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE

Concerted Action in the Framework
of the Copernicus Program



Newsletter 6 *August 1997*

Contents

1. Editorial ---------------------------- 3
2. TELRI Event - Kaunas 2nd TELRI seminar ------ 4
3. TELRI Event - Birmingham workshop ---------- 10
4. On the lexicons (continued) --------------- 18
5. New prospective member of the TELRI advisory board - 33
6. List of Participants -------------------- 35

Coordinator:

Dr. Wolfgang Teubert
Institut für deutsche Sprache
P. O. Box 101621
D - 68016 Mannheim, Germany
phone: +49 621 1581 437
fax: +49 621 1581 415
e-mail: telri@ids-mannheim.de

Editors:

Prof. Eva Hajičová
Mgr. Barbora Hladká
Institute of Applied and Formal Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25
118 00 Prague 1, Czech Republic
tel.: +42 2 21 91 42 88
fax: +42 2 21 91 43 09
e-mail:
hajicova@ufal.mff.cuni.cz
hladka@ufal.mff.cuni.cz

Contributions to
TELRI newsletter,
and address corrections,
should be sent to:

e-mail
hladka@ufal.mff.cuni.cz
fax: +42 2 21 91 43 09

Editorial

Ruta Marcinkeviciene, *Vytautas Magnus University*

It is rather symbolic that the second TELRI seminar „Language Applications for a Multilingual Europe“ took place in Lithuania, Kaunas, not far from Europe's geographic centre in the country which is between the East and West, having emerged from the Eastern block and seeking its place in the European Union. Besides, Lithuania was a suitable location for the seminar because as a small country speaking one of the most ancient Indo-European languages it is aware the importance of the issues of communication and language preservation.

Before the seminar there was a plenary meeting of TELRI, which is a three year project. This being the final year of the project allowed us to retrospect and evaluate the previously achieved results and the unaccomplished tasks in separate work groups. It was possible to look back and recallthe beginnings of TELRI which seemed so far away and to proclaim with John Sinclair „*We all went a long way*“. From a shy beginning discussions about what kind of computers we work with to more bold collaborative endeavours culminating in proposals for new projects. These proposals - TELPROM, PAROLE East and TELRI Association - indicate that TELRI has achieved its goal to create trans-European language resources infrastructure. We will reap the fruits of our labour in the future, but the participation in the project itself is immeasurably valuable, especially for newcomers to the field. This seminar, having gathered more than 70 participants from 25 countries, was remarkable for dr. Wolfgang Teubert's well balanced program (industry versus academy, reports versus demos, theory versus practice) and the participants´ genuine interest in currently actual issues such as standardisation.

Discussions, which began with Jeremy Clear's report, continued during the break and went into the farewell party. Lithuanian folk songs started this party and turned into multilingual singing directed by Antonio Zampolli. This general merriment displayed itself by various activities such as looking for night life in Kaunas and accidental exchange of jackets. This could have had sad results if not for wearing name-tags which indicated real ownership. Although the seminar concluded there still remain discussions to finish, topics to explore and projects to be worked on. Hopefully this will be accomplished during the 3rd TELRI seminar in Italy in October. Wishing the organisers energy and success I add: „*Don't forget name-tags*“.

TELRI[1] event:
Kaunas - 2nd TELRI seminar

Industry and Academia: The Turning Point

Uri Zernik,
OpenSource Inc., New York

Somewhere in the last decade, a role reversal has taken place in our professional world. Traditionally, Academia (mostly the exclusive club of the ten leading American universities plus the major research labs), enjoyed a tremendous lead in what we call the „hi-tech" fields.

However today, the picture is totally different. With the advent of effective communications over the Internet, and the free access to vast text resources on-line, knowledge has disseminated across the board. New text-processing products are being developed along with new experimental techniques in companies such as Netscape, Yahoo, Microsoft, Lexis, and many other smaller companies.

We now live in an egalitarian world. We all have a chance to contribute to the game, and to be active players in the information marketplace. All one needs is some solid natural processing technique, a PC, and a hook-up to the Internet.

In my talk I will describe these trends based on my own personal experience. I will chart some ways in which we computational linguists can play our role in the global game.

A ‚Hopeless' Project:
The English-Slovak & Slovak-English Dictionary (ESSE)

Vladimír Benko
Comenius University, Bratislava
e-mail: jazybenk@savba.savba.sk

The presentation describes a joint venture between a commercial publishing house (Slovak Pedagogical Publishers) and our Laboratory, in the framework of which a methodology has been developed for tagging of a dictionary text that had been (by a rather incompetent decision during the previous stage of the project) originally keyboarded as a ,plain' text without any mark-up. (At the moment of our joining it, the Project, described as ,hopeless' by the publisher, had been going on for more than ten years already.)

---

[1] http://www.ids-mannheim.de/telri/telri.html

After analysing the data, a simple set of software tools has been designed providing for semi-automatic assignment of the dictionary entry structure tags. Based on regular grammar rules, the system is able to recognise the headwords, morphological information, grammatical and stylistic labels, and sense numbers. Moreover, it is able (to a certain extent) to tag the English and Slovak parts of the example phrases. The implementation of the system is based on simple tools (mostly written in lex) and statistics performed in a ,bootstrap' way.

Morphological Analyzer

Svetlana Stoyanova-Goranova,
Bulgarian Academy of Sciences, Sofia
e-mail: jpen@bgearn.bitnet

A morphological analyzer for the recognition of the word forms of Bulgarian has been developed in the Department for Computer Modelling of Bulgarian in the last two years. It is based on the system "Plain" of Prof. Peter Hellwig and programmer on Turbo-C. The linguistic database represents a base lexicon that comprises a dictionary of stems and a dictionary of inflections. The former includes 3000 units from all parts of speech. The latter is a basis for the analyses of all wordforms of the inflectional paradigms of the stems. In the cases of homonymy all and only the right analyses are obtained. This year an extension of the dictionary of stems has been begun.

A further step is lemmatizer which can be used to define the lemmas of the words, but besides lemmatization it also provides explicit morphological analyses and automatic updating of the dictionary of stems. As a rule, the lemmatization is performed automatically without the participation of an operator. Only in the cases of homonymy of word forms - when a wordform has more than one analysis and more than one lemma - the operator selects the appropriate lemma from the variants offered while the rest is erased.

The module for automatic updating of the dictionary is activated when the text contains a wordform which cannot be recognized by the morphological analyzer. Then the operator has to insert the stem(s) of the inflec-

5

tional paradigm, to determine the part of speech and to select from the offered menu the inflections which each of the new stems take. The menu does not contain the inflections for the formation of all the forms of the inflection paradigm, but only those of the forms necessary and sufficicent for the assignment of each of the new stems to a given inflectional type, which is done and included into the database automatically. This makes possible the recognition and lemmatization of any wordform of the paradigm at its next occurrence in the text.

The product can be used for automatic lemmatization of texts of arbitrary length, including corpora, replacing the manual lemmatization which depends entirely on the qualification of the operator and which is often realized with mistakes, especially in the case of homonymy.

Parallel Corpora and Equivalents not Found in Bilingual Dictionaries: An Attempt at Their Generalisation

Alexandra Jarošová,
Slovak Academy of Sciences, Bratislava
e-mail:sasaj@juls.savba.sk

An analysis of the English and Slovak translations of Plato's Republic was done with respect to the three groups of translation equivalents (TEs) proposed by W. Teubert:

A. TEs found in bilingual dictionary (BD)
B. TEs not found in BD and not regarded as suitable
C. TEs not found in BD but regarded as suitable

The results of the A list analysis were presented at the TELRI seminar in Ljubljana (1997) and published in TELRI Newsletter No 5 (April 1997). The C list contains, as a rule, context-sensitive but recurrent TEs.

The following types of TEs are missing in the English-Slovak Dictionary:

1. Slovak verbs as equivalents of English noun-verb collocations. The absence of these equivalents in a given English noun entry is the result of insufficient treatment of collocations consisting of the noun (headword) and verbs „denoting creation and activation" (BBI Combinatory Dictionary's expression), e.g. offer up prayers to sb., to catch sight of st./sb., take hold of st./sb.

2. Slovak adverbs as equivalents of English prepositional phrases. The absence of these equivalents in a given English noun entry is caused by

inconsistency in presenting the noun as a part of prepositional phrases, e.g. at present, in justice, from a distance.

3. Synonyms of prototypical equivalents as translational devices applicable in specific contexts:

a) The synonym of a prototypical equivalent compensates for the collocational restrictions of the latter.

b) Despite no restrictions on the use of the prototypical equivalent in a given lexical environment the synonym of the canonical equivalent has a more specialised meaning for a given lexical partner

Two related issues will be considered:

What is the nature of the prototypical equivalent in existing bilingual dictionaries?

The lexicosyntactical environment of the lemma and the problem of arrangement of BD entry structure.

CUE – A Software System for Corpus Analysis

Oliver Mason,
University of Birmingham, Birmingham
e-mail: O.Mason@bham.ac.uk

Since I might not be able to give a ‚live' demo (depending on the accessibility of a suitable machine) I have prepared a more general presentation. This will start off by explaining the qualitative differences in handling between small and large corpora and the problems that one faces when working with large corpora. Then the solutions adopted in CUE will be explained and its main features described. If possible a ‚hands-on' demonstration follows.

MARK ALISTeR: Marking, Aligning and Searching Translation Equivalents

Stoyan Mihov,
Bulgarian Academy of Sciences, Sofia
e-mail: stoyan@lml.acad.bg

MARK ALISTeR is a system for automatic aligning and searching of translation equivalents in large bilingual corpora. It performs sentence alignment of parallel texts using the Gale-Church algorithm, with resulting correctness of more than 95%.

MARK ALISTeR accepts input files of different formats: .txt files (with or without line breaks), WinWord files, files with Ventura markers, SGML marked text (with or without sentence marking). The correctness of marking is checked as well. The automatic searching and the display of translation equivalents under the synchronised flow of the parallel texts is another extremely helpful function of MARK ALISTeR.

MARK ALISTeR was developed at the Linguistic Modelling Laboratory, Bulgarian Academy of Sciences. It is an MS Windows application running on all INTEL-based Windows systems after Version 3.1 with user interface written in DELPHI. The system was designed and implemented in order to facilitate our tasks of aligning corpora in GLOSSER #343 COPERNICUS'94 JRP and BILEDITA #790 COPERNICUS'94 JRP.

In its actual version MARK ALISTeR is a language independent tool. The quality of the alignment results of MARK ALISTeR can be improved by:
(1) decreasing the noise level in the input data by integration of language specific information for correct recognition of sentence boundaries;
(2) elaboration of the editing facilities of the system.

Czech lexicon by two-level morphology

Hana Skoumalová,
Charles University, Prague
e-mail: Hana.Skoumalova@ff.cuni.cz

In my paper I show the way how I converted an existing Czech lexicon to a two-level morphology system. The existing lexicon was originally designed for simple C programs that only attach ‚endings‘ or ‚suffixes‘ to ‚stems‘. The quotation marks in the previous sentence mean that the terms stem, ending and suffix are used technically rather than linguistically. All alternations were handled inside the endings or suffixes, which required to create more paradigms than really exist in the language.

In the two-level approach, it is possible to work with a morphonological level and then to treat the phonological and/or orthographical changes by separate rules - two-level rules. In our lexicon I first had to redesign the set of paradigms. Those paradigms that only differed in the phonological alternations were merged and they were rewritten from the orthographical form to morphonological form.

The next step was to create a set of two-level rules. In my work I did not try to cover all alternations that occur in the language, but only those that are frequent and productive. Other alternations are either treated as exceptions (e.g. shortening the vowel in a noun stem) or several stems are introduced for one lemma (e.g. six stems for irregular verbs - for infinitive, present indicative, imperative, past participle, present participle and transgressive). The three main types of alternations covered by the two-level rules are palatalization, assimilation and epenthesis.

The whole lexicon that has been converted contains about 17 million word forms representing about 35 million grammatical forms, which covers about 96% of a running text.

TELRI[1] event: Birmingham workshop

Kiril Ribarov, *Czech Republic*
e-mail: ribarov@ufal.mff.cuni.cz

In TELRI Newsletter 3, June 1996, one could read a very enthusiastic and promising report on the first TELRI workshop held October 10-13, 1995, in Birmingham. The latest event, **Birmingham revisited** (May 26-30, 1997), has successfully demonstrated the continuation of the corpus oriented activities and it has fulfilled the expectations one could hold after reading the first report; it has also brought additional strength and encouragement. The workshop took place at the School of English at the University of Birmingham, situated in the red-brick heart of the University, close to the gracious University Tower (which very much resembles of a classical Italian one) and the University Library.

There were sixteen participants present: Janusz Bien and Zygmunt Saloni from Warsaw, Kadri Vider from Tartu, Aneta Dineva and Iordan Penčev from Sofia, Andrejs Spektors from Riga, Alexandra Jarošová form Bratislava, Tamas Varadi from Budapest, Tomaž Erjavec from Ljubljana and František Čermák, Karel Kučera, Jan Hajič, Zdeňka Urešová, Jaroslava Hlaváčová, Alena Böhmová and me from Prague. Many of us were young researchers and students.

The four day workshop enriched us in many ways: carefully selected lectures, open dialogues, new experience and all of it accompanied by frank and warm hospitality of our organiser, **Ann Lawson**, with local help from **James Williams** during the event itself. And guided by English habits, all of us enjoyed warm sunny weather. I will allow myself to say, and I hope that the participants will agree, that the organiser arranged everything we could think of, starting with smileful invitation at the place of our arrival, accommodation arrangements with the rich English breakfast which everybody got used to very fast, access to the Internet, access to the university facilities including the University Library, lot of information concerning the cultural events, the university campus, the town of Birmingham and, of course, the Cadbury chocolate factory. Writing about chocolates, it reminds me of two things: Roald Dahl and his book „Charlie and the Chocolate Factory", and the pleasant and very rich dinners all of us experienced in a French and several Indian restaurants, again thanks to our organisers. Also, our conversations continued many times after the lectures in a tempting atmosphere of the English pubs.

---

[1] http://www.ids-mannheim.de/telri/telri.html

In the following I will devote myself to the lectures. After the workshop opening by Ann Lawson on Monday, the workshop started with the lecture **„Demonstration of Bank of English and tools for collocational analysis"** by **Geoff Barnbrook** who talked about the historical dimension of corpus research using the up to date Cobuild tools. Cobuild[2], which we visited on Wednesday, was present in all workshop lectures: it prepares, preserves and analyses the evidence of the English language - corpora of the English language, the Bank of English, structured into subcorpora (includes British, American and Australian English) with currently more than 323 million nodes. The nodes[3] (words) have their POS tags and are lemmatised. Most of the functions one needs to analyse the behaviour of the nodes within the structure of the Bank of English are gathered under the modules of XLOOKUP, as the starting point for collocational analysis, the tenet of which is, as stated by the speaker, that words do not occur accidentally or randomly, but that there are constraints which cause some of the words to be more in the vicinity of the other words. With the XLOOKUP tools it is possible to access nodes or group of nodes (by defining a simplified regular expression). This output could be declared to be an autonomous unit, a subcorpus, allowing us to perform further analysis with respect to it. An interesting feature is the so-called picture of a located node, which is presented as a table computed from collocations. The picture shows the most frequent words which happen to occur on the first, second, third etc. positions on the both sides of the node, mutually unrelated and ordered by the frequency of occurrence. This is a useful tool which summarises the information from, very often, long list of collocates. There are also tools, which help one to obtain pure statistical information, as the t-score, which offers a comparison of a random occurrence of a node with its actual occurrence in the corpus, or other scores based on the mutual information principle. The statistical measures may serve as criteria for sorting. We were shown how the corpus and the XLOOKUP tools may be used, e.g. in order to trace the history of spelling of English, the result of which is that spelling changes are registered such that irrational spelling is forwarded. It is known that the 16th century introduced more complicated spelling, which was the time when English had become more self confident. It was thought that if the spelling were more difficult (irrational) it would be more respectable. Rather provocative might be the speaker's statement which the speaker stated at the beginning, which is that he sees no differences between literature and linguistics.

---

[2] http://titania.cobuild.collins.co.uk/

[3] The tokeniser separates the tokens in a ‚hard' way. Eg. the abbreviated „won't" is tokenised in to „won" and „t" (two nodes). This introduces unpreciseness, but it was claimed to have more advantages than disadvantages. That is why the term node is preferred.

The second talk **„Standards for Data Reference“** was delivered by the host of the previous Birmingham Workshop, the father of the corpus research at the Birmingham University, **Prof. John Sinclair**. He focused on very crucial, basic but extremely important questions concerning the preserving and representation of information encoded in natural language. One could only agree, that the TEI standards are too high to be acceptable, that there is a need of meta characters over ordinary characters. He pointed out 3 modes in which the language is acceptable: spoken (being linear), written (being non linear) and electronic (linear), which he assumes as a different mode because it is already electronically encoded. In the further processing of any texts, it is very important to preserve the textual integrity of the document, which could be obtained by preserving the original while making its digitised copy. Sinclair distinguishes the digitised copy from the so called working copy[4] (a subset of the digitised copy), which is a single transcription of the digitised copy. Any further processing is done on the working copy only, which is being aligned to the digitised copy by an aligner. This approach avoids the obstacles of the current practise: it is based on single linear string, it keeps to the tradition of manual mark-up, it has a lack of distinction subjective/objective and it tends to impose document models (that the document should look as defined by the style, so the author cannot change the document without changing the style first). The general approach, which Sinclair tries to build, is assumed to be an automatic, on-line and multiple stream data processing.

As stated earlier, Cobuild was „everywhere“- a reception took place at the Westmere place, the former Cobuild site, yet a place full of literature, a place where Shakespeare plays are being performed in its beautiful gardens. Everybody was already well-acquainted with the Workshop environment and the reception was a very suitable place for exchanging our first impressions and getting more in contact with each other.

We also met some of the people from **Cobuild**, when the Workshop continued during Wednesday morning, at their new site on the other side of the

---

[4] The working copy may include other characters, but is very close to the so called zero level copy (bare alphanumeric text, nothing else is there): legibility, alphanumeric text (including punctuation, CRLF).

Campus with a rich and nicely organised program. After the opening talk by Jeremy Clear, we heard about **„Corpus-based English grammar analysis"** by **Gill Francis**. She explained to us the way how the grammar is described in the Cobuild dictionaries, namely the verb frames. After that **Ramesh Krishnamurthy** talked about **„The Bank of English"**. In an open discussion we were acquainted with the types of materials included in the corpus and the difficulties and costs involved in collecting materials for the corpus, form 60 GBPs for electronically available materials up to 15000 GBPs for manual insertion of texts available on audio tapes (for 1 million words). In order to „weight" the corpus with a variety of types of texts, they collect a so called ephemera, a collection of newspaper and magazine headers, advertisements, which are being manually typed.

We were given the opportunity to have a direct look at how the lexicographers compile the dictionaries and how they use the XLOOKUP tools in order to analyse a dictionary entry; Ros Combley, Jenny Watson, Laura Wedgeworth and John Williams were of a great help and showed significance patience.

Then, our visit continued in three parallel sessions: **„Phraseology"** by **Rosamund Moon**, **„Dictionary project management"** by **Stephen Bullon** and **„Computational aspects of Cobuild work"** by **Jeremy Clear**. I owe my thanks to my colleagues, who were there, and explained to me what was happening on the other two parallel sessions and allowed me to use their materials here in this report.

R. Moon talked about some possible approaches to corpus analysis for meaning extraction, pointing out the difficulties when one tries to do so with a wide spectrum of warnings. She said out that in a large corpus it is necessary to define accurately combinations of words, mostly based on their experience from their work on definitions of idioms and phrases on the bases of collocational patterning (1987); this experience was negative, since it was very difficult to locate the boundaries of the collocations. This is an area with strong diversities and dependence on style (formal, conversation, fiction, non-fiction). Idioms were introduced as a special kind of phrases. They tried to capture their frequency in usage, to answer the questions of their heritage, being aware of new American idioms in British English (mostly in journalism), and different metaphor shifts (health metaphors in financial text) etc. Also, other varieties are involved since patterning in spoken English is different than that of written English. The research on meaning extraction is very close to the research of locating of idioms. In this area some statistics were reconstructed: counting frequencies of groups of immediate neighbours, or frequencies of literal versus idiomatic meanings in the Bank of English. When asked why English is so idiomatic, she answered: „We develop new concepts - so we need new expressions for them, but instead of creating brand new words, we use already existing words and put them together in order to bring new senses. Evidence for this can be found in the corpora."

The second parallel session, „Dictionary project management", was directed by Stephen Bullon, who attracted our attention to the more practical side of how to combine the dictionary development with the current market conditions. On the one hand, all of us heard what we had already somehow experienced, but on the other hand, it was a kind of a relief in confirming the actual marketing problems: the financial and marketing situation often and almost always influences the final product in many ways.

Jeremy Clear is the author of the XLOOKUP system and he talked about the „Computational aspects of Cobuild work". In his talk he gave a nice survey through corpora, their analysis, compiling a dictionary and its final printing. He presented a more profound background of the done work at Cobuild and documented their slogan: „The bigger, the better!".

In the afternoon, after a short walk to Westmere, we heard two lectures on **„Data-driven learning"**, the first being **„Monolingual and Multilingual Data and Software"** by **Philip King**. The work on multilingual corpora started in 1994, based on the idea of Francine Roussel (Université Nancy II). The project has its current partners in Denmark, Finland, France, Germany, Italy, Spain and the UK, and it incorporates Danish, English, French, German, Greek, Italian (originally since 1994), with Finnish, Portuguese, Spanish, Swedish being officially added in 1997. There is also a group of unofficially added languages, which are: Afrikaans, Dutch, Hungarian, Lithuanian, Polish, Russian, Welsh and Zulu. This very ambitious project plans to incorporate also Chinese, Japanese and possibly other languages in the future. The aim is mostly pedagogical: it should serve language teachers and learners as well as translation trainers and trainees. The pedagogic focus requires: an easy mark-up, an easy input of own text pairs, a user friendly interface and student control, a possibility of test-creation and an effective feedback between programmers and users. We were given the chance to get fully acquainted with this software. The future development will include more languages, more text(s) types, more pooling of experience, greater interactivity and more local autonomy.

After a short break **Tim Jones** continued this afternoon with a talk on **„Monolingual and Multilingual Teaching Materials"**. The roots of the methods of Corpus Linguistics go back to 1960's. By the way, some of the participants recalled Tim Jones as their English teacher at their local universities! Even now, he permanently helps foreign students with English. He guides himself by the metaphor of a learner being a researcher, testing hypothesis and revising them in the light of data - or as a detective, finding and interpreting linguistic clues. Data Driven Learning changes perception not only of how to organise learning, but also of what is to be learned. We enjoyed his approach and methods, which he illustrated by rich lists of examples.

An English breakfast and a sunny weather opened a new day in Birmingham, which everybody was looking forward to. **„English Words in Use - Compiling a Dictionary of Collocations"** by **Ann Lawson** was the first Thursday lecture. She presented a part of the work she undertook while still based in Birmingham, before moving to work at the IDS in Mannheim, bringing us a new untraditional dictionary of collocations instead of, I would say, limited explanations. As stated by A. Lawson, collocations are hard for standard dictionary to describe, since they are flexible, discontinuous, introspective and intuition inaccurate. So, from the learners point of view, they are opaque and tricky, require experience and they account for many mistakes. Also, everyday and frequent collocations are very easy to miss and thus difficult to find. One of the basic aims of this dictionary is thus to catch those kinds of collocates. The dictionary is supposed to be finished at the end of 1997. Let me express my wish for a success of this novel approach.

Originality of approaches persisted through the next lecture by **Oliver Mason**, **„Lexical Gravity"**. He stated that collocations, in the present works, have not been parametrised. He tried to do so, by defining the following collocation parameters: environment span, cut-off/threshold (throw away words with freq < n, where n is small - they are either misspellings or very rare words - it depends on the choice of n), node preprocessing - groupings (semantical, uppercase, lowercase etc.), collocate preprocessing, significance evaluation (mutual information, t-score etc.) and reference frequency. Further he specified a context (span) as something which defines a specialised sub-sample (sub-corpus) and is motivated by syntax (sentence, phrase), distance (window), adjacency and has influence on result and computational costs.

After the definitions, the author concentrated on the formalism concerning the measuring of the influence of nodes on each other. To do so, he accepts the following assumptions/predictions: the variability of environment is influenced by the node; there are different patterns for grammatical and lexical items; there are individual patterns of influence for each word; there exists an upper limit on the span of influence; the results should be independent of the taken sample. The procedure for measuring the mutual influence calculates the TTR (true token ratio) for each relative position of a node word, after a collection of its instances has been done. The result of this procedure should be a certain threshold of significance. The graphical interpretation very much reminds of a gravitation gap, thus lexical gravity. Oliver Mason's analysis has come to the following conclusions: lexical gravity is not symmetrical, different words have different patterns, different forms have different patterns, there exists a constant over different corpora, the lexical gravity is more stable with an increasing size of the corpora and an existence of certain, so called ‚negative' gravity for certain grammatical words was postulated.

His future plans include classification of words according to their gravity patterns, separation of different meanings of forms; he plans to take into consideration multiword expressions and fixed phrases and to investigate of how other languages behave in the described formal sense.

From my experience, I would like to add, that if similar results are run on letters instead of words, a certain isomorphism could be observed. The experiments are even more encouraging, since something similar to lexical gravity could be reconstructed from autocorellation function when run on either letters or nodes.

A factual prove on what has been said so far, illustrated by a plenty of examples was provided by **Prof. Frank Knowles** in the next session, **„Corpus Analysis for LSP"**, where the lecturer was trying to explain the vagueness of the words when compared to their forms (the words themselves).

The last day of the Workshop was a day of an open dialogue, and a day where the participants had the opportunity to present their own work. **Jan Hajič** from Prague presented the **Czech National Corpus** and **Prague Dependency Treebank** (see TELRI Newsletter 4 and 5, for more details about both of these projects).

The last presentation was about the **„WordSmith Tools"**[5], by **Mike Scott**. WordSmith Tools is a package of programmes that help to see how words behave in texts: the Wordlist tool provides a list of all words or word-clusters in a text, set out in alphabetical or frequency order; Concord, a corcondancer, gives one a chance to see any word or phrase in context; KeyWords finds the key words in a text.

We were a group, on one hand, big enough to raise issues and contrast ideas, and on the other hand, small enough to cooperate and to sit at one table, let's say in a nice English pub or restaurant. For some of us, it was the first time to visit Birmingham and the University Campus and we brought back wonderful memories. For the younger of us, it was encouraging to meet other young researchers from the workshop group, as well as from the host university itself.

---

[5] http://www.liv.ac.uk/~ms2928/wordsmit.htm

The weather was still sunny, even the day we had to leave.

„It was a pleasure to welcome on behalf of TELRI a mixed and interesting group of researchers to the four-day workshop. The enthusiasm, not to say stamina, of the participants, together with a varied programme, made for a very enjoyable and productive time. I especially enjoyed the Czech presentation as an example of real bilateral participation in the workshop. An additional side-effect was that the visitors' requests prompted me into discovering that it is possible to ascend the clock tower in the centre of campus, which we promptly did. During almost ten years at the University I had never done this and, surprisingly enough, the English weather smiled on us to grant us wonderful views. In summary, a very worthwhile and thought-provoking time was had by all."

Ann Lawson, IDS, Mannheim

On the Lexicons

(continued)

Verb frames in the Czech hierarchical lexicon

Hana Skoumalová
Institute of Theoretical and Computational Linguistics
Charles University, Prague
e-mail: hana.skoumalova@ff.cuni.cz

ABSTRACT

In the Czech hierarchical lexicon which I created, the main stress was put on the verb frames. In my paper I want to describe the main classes of verbs in Czech and the way in which I treat them in the lexicon. I will also discuss some interesting theoretical problems connected with verbs.

First I will make a brief introduction to the valency theory that is the backbone of my description of the verb frame format. Further, I will describe and demonstrate on examples the base format of the verb frame in my notation. In the next section I will discuss the relationship between the active and the passive frame, and then I will study properties of equi and raising verbs.

1 THEORETICAL BACKGROUND

Though I want to build the lexicon as theory independent, in the background *some* theory must be present. At least, the lexicon must include basic linguistic categories. The only requirement is, that the notation can be converted to another notation (e.g., HPSG or Dependency Syntax).

As the "background" theory I utilized the theory developed by Sgall, Hajičová and Panevová [SHP86], and especially the part dealing with the verb frames [PAN74], [PAN75], [PS92]. Two levels of syntactic description–the underlying (e.i. deep) structure and the surface structure–are distinguished. In the underlying structure we work with inner participants (*actants*) and free modifications. Each verb can have up to five inner participants: Actor, Patient, Addressee, Origin and Effect. These inner participants are mem-

---

[*] This work was partly supported by grant No. 72\94 of Research Support Scheme

bers of the verb frame and they are realized as the subject and objects in the surface structure. Some of them can be *optional*, which means that they do not need to be present in the sentence–both in the underlying and surface structure. Other participants are *obligatory* in both structures, and another sort of participants are those that are obligatory in the underlying structure but can be omitted in the surface structure. These are called *obligatory deletable* participants. Whether a participant is optional or obligatory deletable can be tested by a question test. Let us imagine the following dialogue:

(1) *–Petr čte. –Co? –Nevím.*
–Petr is reading. –What? –I don't know.

The answer 'I don't know' is acceptable, as the the speaker does not need to know what Petr's reading is. This shows that Patient of the verb *číst* (to read) is optional. On the other hand, in the following dialogue, the sentence with deleted Actor is acceptable but the answer 'I don't know' is nonsensical. This shows us that the Actor is an obligatory deletable participant.

(2) *–Už přišel. –Kdo? –*Nevím.*
–(He) has already come. –Who? –*I don't know.

In the next example the sentence is actually ungrammatical, if the participant is omitted–this is a clear evidence that the participant is obligatory:

(3) **Petr daroval.*
*Petr donated.

Beside the inner participants, the verb frame may have also other members–the adverbials (adjuncts, free modifications)–but only if they are obligatory in the underlying structure. However, they can be deletable on the surface:

(4) *–Petr přišel. –Kam? –*Nevím.*
–Petr came. –Where? –*I don't know.

2 VERB FRAME

The formalism I used for the lexicon is DATR–a formalism designed for creating hierarchical structures, and especially dictionaries. For more information about its properties see [Gaz90].

The whole structure of the lexicon consists of two, non-overlapping parts: a morphological and a syntactic one. In this article I will mainly deal with the syntactic part.

The lexical entry has the following form:

```
L_bát_1:        <> == VERB                                        % BE AFRAID
                <gloss> == 'bojí se strašidel;
                                      ... že nepřijdou;
                                      ... aby nepřišli'
                <mor> == BÁT
                <syn> == RSE_F[2|clz|cla]@.
```

The base form serves as a node in the hierarchical structure (it starts with L_ and can be followed by an index _1 that distinguishes different lexical meanings of an ambiguous lexical item). Every entry contains gloss with an explanation of the meaning or examples of the usage, and mor and syn part. The verb in the example inherits default values, which can be overwritten, from the node VERB, all the morphological information from the node BÁT and the syntactic information from the node RSE_F[2|clz|cla]@. After querying the system we get this output (only the syntactic information and the gloss is shown):

```
L_bát_1:        <gloss> = bojí se strašidel;
                                      ... že nepřijdou;
                                      ... aby nepřišli
                <syn cat> = V
                <syn type> = main
                <syn refl> = se
                <syn subj surf> = NPnom
                <syn subj deep> = Actor
                <syn subj oblig> = oblig_deletable
                <syn 1_obj surf> = NPgen , CLže , CLaby
                <syn 1_obj deep> = Patient
                <syn 1_obj oblig> = optional
                <syn pass> = no.
```

The syntactic category of the word is V, its syntactic type is main verb, and the verb is intrinsic reflexive (*reflexive tantum*), with the reflexive particle se. The frame has two members: Actor and Patient. Actor plays the role of Subject, it is obligatory deletable, and in the surface structure, it occurs as a noun phrase in Nominative. Patient is optional, plays the role of the first object and in the surface structure it can occur as an NP in Genitive, or as a clause, connected either by the conjunction *že* or by the conjunction *aby*. A passive construction of this verb is impossible.

The syntactic part of the structure 'above' the node L_bát_1 is shown in the next example:

```
SIGN:          <> == UNDEF
               <gloss> == .

VERB:          <> == SIGN
               <syn cat> == 'V'
               <syn type> == main
               <syn refl> == no
               <mor infl neg> == (ne "<mor infl>").

SD1:           <syn subj surf> == 'NPnom'
               <syn subj deep> == 'Actor'
               <syn subj oblig> == oblig_deletable.

D_O4:          <syn 1_obj surf> == 'NPacc'
               <syn 1_obj deep> == 'Patient'
               <syn 1_obj oblig> == obligatory.

D_O2:          <syn 1_obj> == D_O4
               <syn 1_obj surf> == 'NPgen'.

D_F2:          <syn 1_obj> == D_O2
               <syn 1_obj oblig> == optional.

D_F[2|clz|cla]: <syn 1_obj> == D_F2
               <syn 1_obj surf> == 'NPgen , CLže , CLaby'.

RD1$:          <> == VERB
               <syn subj> == SD1
               <syn pass> == refl.

RSE@:          <> == RD1$
               <syn refl> == se
               <syn pass> == no.
```

```
RSE_F[2|clz|cla]@:  <> == RSE@
                    <syn 1_obj> == D_F[2|clz|cla].
```

The highest node in the hierarchy SIGN does not assign any values to any attribute. The node VERB assigns the value V to syntactic category; syntactic type is assigned the value main; a verb is by default non-reflexive and negative forms are created by a prefix *ne-*. In the above hierarchy also three frames are defined: RD1$ defines an intransitive verb, RSE@ defines an intrinsic reflexive verb, and RSE_F[2|clz|cla]@ defines an intrinsic reflexive verb with an object in Genitive or in the form of a clause. The definitions of frames are constructed from definitions of the subject (SD1) and the objects (D_O4, D_O2, D_F2 and D_F[2|clz|cla]). Every frame also determines whether a verb can form the passive voice, and what sort of passive voice is appropriate.

3 *PASSIVE VOICE*

Up to now I have been only speaking about active frames of verbs. The question is, whether we are able to derive the passive frames from the active ones, or whether the passive constructions must be stored separately in the lexicon as well.

In Czech two kinds of passive exist:

periphrastic uses the auxiliary verb *být* (to be) and a passive participle
kniha je čtena–the book is read

reflexive (mediopassive or impersonal passive) uses a finite form of the verb and the reflexive particle *se*
kniha se čte–the book SELF reads
jde se – it goes SELF

As the passive is created in a regular way, the information about the kind of passive is sufficient for us. Passive constructions can be created by lexical rules or by rules of the grammar according to the following algorithm:

- If 1st object in the frame is in Accusative, it becomes the subject (in Nominative).

(5) *Čtu knihu*$_{Acc}$. – *Kniha*$_{Nom}$ *je čtena*. – *Kniha*$_{Nom}$ *se čte*.
I read the book. – The book is read. – The book SELF reads.

- If 1st object is a clause or an infinitive, it becomes the subject, with a special sort of agreement (3rd person, singular, neuter).

(6) *Dokážu, [že je to pravda]*$_{Acc}$ & – *Bude dokázáno, [že je to pravda]*$_{Nom}$.
I will prove that it is true. & – It will be proven that it is true.
– *Dokáže se, [že je to pravda]*$_{Nom}$
– It will prove SELF that it is true.

- If the 1st object has a form different from those quoted above, or is missing, the passive has empty subject, with the same sort of agreement as the infinitive or clause subject.

(7) *Dosáhneme vrcholu*$_{Gen}$ *hory.*
We will reach (of) the top of the mountain.

– *Bude dosaženo vrcholu*$_{Gen}$ *hory.*
– It will be reached of the top of the mountain.

– *Dosáhne se vrcholu*$_{Gen}$ *hory.*
– It will reach SELF of the top of the mountain.

- The original subject becomes a complementation in Instrumental (for periphrastic passive).

(8) *Kniha byla napsána slavným autorem*$_{Ins}$.
The book was written by a famous author.

- All other members of the frame stay in their positions.

The question remains, what happens with the subject in reflexive passive. It seems that the subject can turn into a complementation in Dative, which sometimes requires another complementation–a modification of manner–to be present in the sentence:

(9) a. *Chci spát. – Chce se mi*$_{Dat}$ *spát.*
I want to sleep. – It wants SELF to me to sleep.
b. *Peču v troubě. – V troubě se mi*$_{Dat}$ *dobře peče.*
I bake in an oven. – In the oven SELF to me well bakes.

The question is whether the complementation in Dative is equal to the original Actor, which usually plays the role of Subject. According to the Czech linguistic tradition the complementation in Dative is rather a modification of an initiator of action. Thus the reflexive passive construction in PS92 will be considered a construction without Actor, while the reflexive construction in BdPC93 will be considered a separate entry in the lexicon, similar to constructions *líbit se komu* ('to appeal to sb') or *zdát se komu* ('to seem to sb').

Another question is, what sort of algorithm applies to verbs with two Accusatives in the frame. There are only two such verbs in Czech:

- *stát koho*$_{Acc}$ *co*$_{Acc}$ –to cost sb sth
  This verb does not have the passive, so there is no problem with this verb.

- *učit koho*$_{Acc}$ *co*$_{Acc}$*/čemu*$_{Dat}$ –to teach sb sth
  There are in fact two frames here: One with Accusative and Dative and the other with two Accusatives. The former does not cause any problems, as it can be treated according the above algorithm. For the latter one more special rule holds: Any of the Accusatives can become Subject of the passive construction, but the other one cannot be present in the structure.

4 VERBS WITH THE INFINITIVE IN THEIR FRAMES

In this group of verbs we have to describe not only the frame of the verb, but also the interaction between the higher verb and the lower verb (the infinitive): Which members of the frames they share, whether the infinitive can be passivized and other constraints that hold for the two verbs.

These verbs are usually divided into two subclasses: **equi** and **raising** verbs. In both cases the subject or an object of the infinitive is one of the participants of the higher verb, but the difference between these two sorts of verbs is in the underlying structure:

raising verb the frame of the infinitive does not overlap with the frame of the higher verb, but on the surface, the verb 'raises' the subject of the infinitive as its own subject or object; this participant occurs only once in the underlying structure

equi verb the frame of the infinitive overlaps with the frame of the control verb; the shared participants occur twice in the underlying structure

4.1 Raising verbs

First, we will deal with subject raising verbs, with which the subject of the infinitive becomes the subject of the higher verb. This group of verbs contains the modal and aspectual verbs. Examples:

(10) *Petr*$_i$ *smí [*__$_i$ *odejít].*
Petr may to-leave.

(11) *Začne [pršet].*
Will-start to-rain.

(12) *Petr*$_i$ *musí [*__$_i$ *být pochválen].*
Petr must be praised.

(13) *Musí [*__$_i$ *se zabít] dvě mouchy*$_i$ *jednou ranou.*
Must SELF to-kill two flies by one hit.
'Two flies must be killed by one hit.'

We see in the examples that the two subjects are shared, no matter which voice is used in the infinitive construction. The higher verb, however, cannot be passivized.

In my lexicon subject raising verbs are encoded in the following way:

```
L_muset:      <syn refl> = no
              <syn subj surf> = 1_obj:<syn subj surf>
              <syn subj oblig> = 1_obj:<syn subj oblig>
              <syn 1_obj surf> = VPinf [pass = perif , refl]
              <syn 1_obj oblig> = obligatory
              <syn pass> = no.
```

The description of <syn subj> contains only pointers to the subject of infinitive (on the surface) and the value of <syn subj deep> is undefined.

The infinitive can occur in both passives, it depends only on whether the verb occurring as the infinitive allows for a passive. The higher verb occurs only in the active voice.

The aspectual verbs imply a further constraint on the dependent infinitive: only an imperfective verb may occur in this position. The verb *začít* (to start) is encoded like this:

L_začít_2: <syn cat> = V
<syn type> = main
<syn refl> = no ,
<syn subj surf> = 1_obj:<syn subj surf>
<syn 1_obj surf> = VPinf [pass = perif , refl; aspect = impf]
<syn pass> = no.

Subject-to-object raising verbs are such verbs that have an infinitive in the frame and the subject of this infinitive becomes an object of the higher verb. This group contains the verbs of perception:

(14) *Vidím ho*$_i$ *[*__$_i$ *přicházet]*.
I-see him to-be-coming.
'I see him coming.'

In the lexicon the frame is encoded in the following way:

L_vidět_2: <syn refl> = no
<syn subj surf> = NPnom
<syn subj deep> = Actor
<syn subj oblig> = oblig_deletable
<syn 1_obj surf> = 1_obj:<syn subj surf> [NPnom = ^NPacc]
<syn 1_obj oblig> = obligatory
<syn 2_obj surf> = VPinf [pass = no]
<syn 2_obj surf> = Patient
<syn 2_obj surf> = obligatory
<syn pass> = no.

The description of <syn 1_obj> contains a pointer to the subject of the infinitive, with the constraint on the case: Nominative must be changed to Accusative. It further overwrites the value of <syn 1_obj oblig> to obligatory.

4.2 Equi verbs

The subject and/or some objects of the infinitive are shared by the frame of the control verb; in the underlying structure, these complementations are present twice.

- Subject-control:

(15) ${}^{Nom}_{Act}$ *Oni*$_i$ *mu slíbili* $[_{Act—i}$ *přijít*$]$.
They to-him promised to-come.

- Subject-control with coindexation of objects:

(16) ${}^{Nom}_{Act}$ *Oni*$_i$ ${}^{Dat}_{Addr}$ *mu*$_j$ *slíbili* $[_{Act—i}$ *donést* ${}^{Dat}_{Ben—j}$ *knihu*$]$.
They to-him promised to-bring book.

- Object-control:

(17) *Oni* ${}^{Dat}_{Addr}$ *mu*$_i$ *poručili* $[_{Act—i}$ *přijít*$]$.
They to-him ordered to-come.

- An object of the infinitive is the subject of the control verb:

(18) ${}^{Nom}_{Act}$ *Plot*$_i$ *chce* $[_{Act—}$ *natřít* ${}^{Acc}_{Pat—i}$ $]$.
Fence wants to-paint.

- The structure can be ambiguous–either subject-control or subject-object coindexation:

(19) a. ${}^{Nom}_{Act}$ *Anežka*$_i$ *chce* $[_{Act—i}$ *číst pohádky*$]$.
'Anežka wants to read tales.'
b. ${}^{Nom}_{Act}$ *Anežka*$_i$ *chce* $[_{Act—}$ *číst* ${}^{Dat}_{Ben—i}$ *pohádky*$]$.
'Anežka wants someone to read tales to her.'

The description of the frame of the verb *bát se (něco udělat)* (to be afraid (to do sth)) looks like this:

```
L_bát_2:        <syn refl> = se
                <syn subj surf> = NPnom
                <syn subj deep> = Actor
                <syn subj oblig> = oblig_deletable
                <syn 1_obj surf> = VPinf [subj = ^Actor; pass = perif]
                <syn 1_obj deep> = Patient
                <syn 1_obj oblig> = obligatory
                <syn pass> = no.
```

The description of <syn 1_obj surf> contains the information, that the subject of the infinitive is coindexed with Actor of the control verb. The reason, why I use this cross-referencing between two strata of the linguistic description, is that this relationship between a participant of the control verb and the subject of the infinitive is preserved even in the passive voice of the infinitive and/or the control verb. I will demonstrate this behaviour on the verbs *chtít* (to want) and *povolit* (to allow):

- Two active voices:

(20) ${}^{Nom}_{Act}$*Já*$_i$ *chci [*$_{Act--i}$ *pochválit* $_{Pat}$*Petra]* .
I want to praise Petr.

(21) *Povolili* ${}^{Dat}_{Addr}$*mu*$_i$ *[*$_{Act--i}$ *přijít]*.
They allowed him to come.

- Active–periphrastic passive:

(22) ${}^{Nom}_{Act}$*Petr*$_i$ *chce [*$_{Pat--i}$ *být pochválen]* .
Petr wants to-be praised.

(23) ${}^{Nom}_{Act}$*Anežka*$_i$ *chce [*$_{Addr--i}$ *být poučena o hudbě]* .
Anežka wants to-be instructed in music.

(24) *Povolili* ${}^{Dat}_{Addr}$*mu*$_i$ *[*$_{Pat--i}$ *být zapsán]*.
They allowed him to be enrolled.

- Periphrastic passive–active voice:

(25) *Bylo* ${}^{Dat}_{Addr}$*mu*$_i$ *povoleno [*$_{Act--i}$ *přijít]*.
It was allowed to him to come.

- Mediopassive–active voice:

(26) *Povolilo se* ${}^{Dat}_{Adr}$*mu*$_i$ *[*$_{Act--i}$ *přijít]*.
It was allowed to him to come.

- Two periphrastic passives:

(27) *Bylo* $^{Dat}_{Addr}$ *mu*$_{i}$ *povoleno* $[_{Pat---i}$ *být zapsán].*
It was allowed to him to be enrolled.

■ Mediopassive–periphrastic passive:g

(28) *Povolí se* $^{Dat}_{Addr}$ *mu*$_{i}$ $[_{Pat---i}$ *být zapsán].*
It will be allowed to him to be enrolled.

The frame of the verb *chtít* is as follows:

```
L_chtít_3:    <syn refl> = no
              <syn subj surf> = NPnom
              <syn subj deep> = Actor
              <syn subj oblig> = oblig_deletable
              <syn 1_obj surf> = VPinf [subj = ^Actor; pass = perif , refl].
              <syn 1_obj deep> = Patient
              <syn 1_obj oblig> = obligatory
              <syn pass> = no.
```

The verb *chtít*, however, has two more frames containing the infinitive. In one of them there is a relation between Actor in the governing clause and Patient of the infinitive:

(29) $^{Nom}_{Act}$ *Plot*$_{i}$ *chce* $[_{Act—}$ *natřít* $^{Acc}_{Pat---i}$ $]$.
Fence wants to-paint.
'The fence needs painting.'

(30) $^{Nom}_{Act}$ *Pepík*$_{i}$ *chce* $[_{Act--}$ *nařezat* $^{Dat}_{Pat---i}$ $]$.
Pepík wants to-spank.
'Pepík needs spanking.'

```
L_chtít_4:    <syn refl> = no
              <syn subj surf> = NPnom
              <syn subj deep> = Actor
              <syn subj oblig> = oblig_deletable
              <syn 1_obj surf> = VPinf [Patient = ^Actor; refl = no; pass = no].
              <syn 1_obj deep> = Patient
              <syn 1_obj oblig> = obligatory
              <syn pass> = no.
```

With the other there is a relation between Actor in the main clause and Addressee or Benefactor in the infinitive clause:

(31) ${}^{Nom}_{Act}$ *Anežka*$_{i}$ *chce* [$_{Act\text{--}}$ *podat* ${}^{Dat}_{Addr\text{—}i}$ *knihu*] .
Anežka wants to-pass book.
'Anežka wants someone to pass her the book.'

(32) ${}^{Nom}_{Act}$ *Anežka*$_{i}$ *chce* [$_{Act\text{—}}$ *přečíst* ${}^{Dat}_{Ben\text{—}i}$ *pohádku*] .
Anežka wants to-read tale.
'Anežka wants someone to read her a tale.'

(33) ${}^{Nom}_{Act}$ *Anežka*$_{i}$ *chce* [$_{Act\text{—}}$ *poučit* ${}^{Acc}_{Addr\text{—}i}$ *o hudbě*] .
Anežka wants to-instruct in music.
'Anežka wants someone to instruct her in music.'

```
L_chtit_5:   <syn refl> = no
             <syn subj surf> = NPnom
             <syn subj deep> = Actor
             <syn subj oblig> = oblig_deletable
             <syn 1_obj surf> = VPinf [Addr/Benef = ^Actor; refl = no; pass = no].
             <syn 1_obj deep> = Patient
             <syn 1_obj oblig> = obligatory
             <syn pass> = no.
```

5 CONCLUSION

The main goal of the lexicon is to capture all morphological and syntactic phenomena in Czech that are important for NLP. The information stored in the lexicon cannot be used in an application 'as such' but must be interpreted. This interpretation, however, should not be very difficult–I tried to make the notation as natural and intuitive as possible.

There are, of course, still some open questions to be answered. Some of them are listed here:

- The aspect of the dependent verb–some verbs (e.g., aspectual verbs) require the imperfective aspect of the dependent verb. Others (e.g., modal verbs) are imperfective, but the whole construction with a modal verb as the main verb gets the aspect from the dependent verb. The question is, whether a construction like

(34) ?? *Přestal*$_{aspectual}$ *smět*$_{modal}$ *tu knihu vydat*$_{perfective}$.
'He ceased to be allowed to publish the book.'

is a correct Czech sentence, and if so, how to account for this construction within my notation.

- The aspect of the dependent verb in the frame of the verb *vidět* (to see)– if the main verb *vidět* is in present tense, it requires an imperfective form of the infinitive, while if in past, both aspects of the infinitive are allowed.

(35) *Vidím*$_{present}$ *ho přicházet*$_{imperfective}$.
I see him coming.
?? *Vidím*$_{present}$ *ho přijít*$_{perfective}$. – Possible as *praesens historicum*?
I see him come.
Viděl jsem$_{past}$ *ho přicházet*$_{imperfective}$.
I saw him coming.
Viděl jsem$_{past}$ *ho přijít*$_{perfective}$.
I saw him come.

These problems still wait for a solution.

REFERENCES

[BdPC93] Ted Briscoe, Valeria de Paiva, and Ann Copestake, editors. *Inheritance, Defaults, and the Lexicon*. Cambridge University Press, 1993.

[EG90] Roger Evans and Gerald Gazdar, editors. *The DATR Papers, Volume 1*. Number 139 in CSRP. University of Sussex, Brighton, 1990.

[Gaz90] Gerald Gazdar. An introduction to DATR. In Evans and Gazdar.

[GK89] Miroslav Grepl and Petr Karlík. *Skladba spisovné češtiny (Syntax of Standard Czech)*. SPN, Prague, 2 edition, 1989.

[Oli89] Karel Oliva. *A Parser for Czech Implemented in Systems Q*. Number XVI in Explizite Beschreibung der Sprache und automatische Textbearbeitung. Matematicko–fyzikální fakulta UK, Prague, 1989.

[Pan74] Jarmila Panevová. On verbal frames in functional generative description, part i. *Prague Bulletin of Mathematical Linguistics*, 22:3–40, 1974.

[Pan75] Jarmila Panevová. On verbal frames in functional generative description, part ii. *Prague Bulletin of Mathematical Linguistics*, 23:17–52, 1975.

[Pan80] Jarmila Panevová. *Formy a funkce ve stavbě české věty (Forms and Functions in Syntax of Czech Sentence)*. Number 13 in Studie a práce lingvistické. Academia, Prague, 1980.

[PS92] Jarmila Panevová and Hana Skoumalová. Surface and deep cases. In *Proceedings of COLING*, pages 885–889, Nantes, 1992.

[SHP86] Petr Sgall, Eva Hajičová, and Jarmila Panevová. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reidel Publishing Company, Dordrecht, 1986.

[Svo62] Karel Svoboda. *Infinitiv v současné spisovné češtině (Infinitive in Contemporary Standard Czech)*. Rozpravy ČSAV. Academia, Prague, 1962.

[Š67] Vladimír Šmilauer. *Novočeská skladba (Syntax of Modern Czech)*. Academia, Prague, 1967.

New prospective member of the TELRI advisory board

Prof. Dr. Alexandr Zubov,
Minsk Linguistic University,
Department of Computer Science and Applied Linguistics,
Minsk, Belarus

The Department of Computer Science and Applied Linguistics was established in February 1975. At present the Department consists of 9 lectures, 3 post-graduates, 3 engineers and 10 members of the technical staff. Ten lecturers of other Departments of the University collaborate with the unit.

Current working projects of the Department:

- computational learning theory and MULTIMEDIA programs development (English, French, German, Spanish, Russian, Belarussian);
- formalization of text structure and development of programs for text generation (French: advertisement, tales, proverbs, riddles, technical descriptions; English: technical descriptions, advertisement; Russian: poetry);
- automatic estimation of lexical stock of foreign language textbooks (on base of statistical coefficients);
- computer understanding and development of programs for information comprimation of texts (scientific, technical, social-political texts).

Our text databases:
1. Scientific Russian text on the theme "Linguistic Computer Science" (near 200 000 units)
2. Scientific Belarussian text on the theme "Linguistic Computer Science" (near 60 000 units)
3. English texts of A. Clarke, G. Greene, I. Murdoch, H. Golding (about 300 000 units)

Our lexical resources:
1. English-Russian dictionary on the theme "Computers, numeric control, data processing in computer network, flexible production systems". The dictionary contained 43 500 words, word combinations and abbreviations.
2. German-Russian dictionary on the theme "Computers, informatics and

robot technology". The dictionary contained 40 200 words and word combinations.

3. Russian dictionary on the theme "Computer technology and programming". It contained 200 most frequently used words of the Russian language, 5 000 terminological word forms and 38 000 stems of Russian words.
4. Russian dictionary of poetry. It contains 3 000 words.
5. Frequency lists (on paper) of 6-alpha characters of combinations (190807), (357 504 entries), 5-alpha characters of combinations, 4-alpha characters of combinations (75 045) and 3-alpha characters of combinations (20 355) of Russian texts. The texts included 520 000 alpha characters of belles-lettres, 170 000 alpha characters of texts of jurisprudence and 310 000 alpha characters of scientific and technical texts.
6. Frequency lists (on paper) of English and French words and word combinations on the theme "Specific systems of communication and computers" (text of each language included 200 000 entries).
7. Russian-English, English-Russian, Russian-French, French-Russian dictionaries. Each pair of languages contained 1 110 - 1 500 words and 170 - 290 word combinations in 12 topic areas (on paper).
8. Russian-English , English-Russian, Russian-French, French-Russian, Russian-German, German-Russian, Spanish-Russian, Russian-Spanish, Italian-Russian, Russian-Italian dictionaries. The dictionaries of each pair of languages contained 1 740 words from 19 topic areas (on papers).

National Project: Creation of Belarussian Computer Fund

International Project: TELRI

List of Participants*:

ANDERSEN Poul	*e-mail:*	m764@eurokom.ie
BECI Bahri	*e-mail:*	beci@igjl.tirana.al
BENKO Vladimir	*e-mail:*	jazybenk@savba.savba.sk
BIEN Janusz S.	*e-mail:*	jsbien@plearn.edu.pl
ČERMÁK František	*e-mail:*	frantisek.cermak@ff.cuni.cz
ERJAVEC Tomaž	*e-mail:*	et@cogsci.ed.ac.uk
FISIAK Jacek	*e-mail:*	fisiak.plpuam11.bitnet
GELLERSTAM Martin	*e-mail:*	gellerstam@svenska.gu.se
HAJIČOVÁ Eva,	*e-mail:*	hajicova@ufal.mff.cuni.cz
HLADKÁ Barbora	*e-mail:*	hladka@ufal.mff.cuni.cz
JAKOPIN Primoz	*e-mail:*	primoz.jakopin@uni-lj.si
JAROŠOVÁ Alexandra	*e-mail:*	sasaj@juls.savba.sk
KRUYT Truus	*e-mail:*	kruyt@rulxho.leidenuniv.nl
MARCINKEVIČIENĖ Rūta	*e-mail:*	ruta.marcinkeviciene@vdu.lt
OIM Haldur	*e-mail:*	hoim@psych.ut.ee
PAJZS Júlia	*e-mail:*	pajzs@nytud.hu
PASKALEVA Elena	*e-mail:*	hellen@bgearn.bitnet
PENCHEV Iordan	*e-mail:*	jpen@bgearn.bitnet
LAURENT Romary	*e-mail:*	Laurent.Romary@loria.fr
LAWSON Ann	*e-mail:*	a.e.lawson@bham.ac.uk
SINCLAIR John M.	*e-mail:*	j.sinclair@bham.ac.uk
SPEKTORS Andrejs	*e-mail:*	aspekt@mii.lu.lv
TEUBERT Wolfgang, VOLZ Norbert	*e-mail:*	telri@ids-mannheim.de
TUFIS Dan	*e-mail:*	tufis@roearn.ici.ro
ZAMPOLLI Antonio	*e-mail:*	paula@icnucevm.cnuce.cnr.it

---

*You can see detailed addresses in Newsletter No. 2.

WHAT IS TELRI

The Concerted Action TRANS-EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE (TELRI), is a COPERNICUS project funded by the European Commission. TELRI has a duration of three years (1995-1997). It brings together 22 institutions of 17 European countries (Albania, Germany, Great Britain, Slovakia, Italy, Bulgaria, the Czech Republic, Sweden, Slovenia, Romania, Estonia, France, the Netherlands, Latvia, Lithuania, Poland and Hungary).

TELRI is setting up a permanent network of leading national language and language technology centres in the whole of Europe. It pools existing language resources, corpora, machine-readable dictionaries and lexicons, lexical databases, and generic software tools for the creation, re-use, maintanance, validation , and exploitation of linguistic data. It complements these repositories with newly created multilingual resources, offering a wide range of language data to the NLP community. TELRI is establishing a platform where research and industry meet, exchange resources and engage in product-oriented cooperation.

Links have been established with language centres elsewhere in Europe, with relevant European organizations and ventures, and with focal language institutions in other parts of the world.

FOR INFORMATION.

Inquiries about TELRI may be addressed to: **Dr. Wolfgang Teubert**, Institut für deutsche Sprache, P. O. Box: 101621, 68016 Mannheim, Germany, Phone: +49 621 1581 437, Fax: +49 621 1581 415, e-mail: telri@ids-mannheim.de

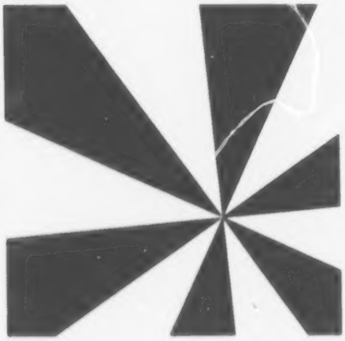
TELRI's WWW Document.

Detailed information about TELRI and its activities is available through the World Wide Web (WWW) at the following URL: **http://www.ids-mannheim.de/telri/telri.html**
Webmaster: Alena Böhmová, e-mail: webadm@smetana.ms.mff.cuni.cz
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Malostranské nám. 25, 118 00 Prague 1, Czech Republic

TELRI

TRANS EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE

Concerted Action in the Framework
of the Copernicus Program



Newsletter 7 *October 1997*

Contents:

1. Editorial ------------------------------ 3
2. Topic of the issue: Multilingual technology ------ 4
3. TELRI Event - Montecatini 3rd TELRI seminar ----17
4. On the TELRI Newsletter -------------------32
5. List of Participants ----------------------34

Coordinator:

Dr. Wolfgang Teubert
Institut für deutsche Sprache
P. O. Box 101621
D - 68016 Mannheim, Germany
phone: +49 621 1581 437
fax: +49 621 1581 415
e-mail: telri@ids-mannheim.de

Editors:

Prof. Eva Hajičová
Mgr. Barbora Hladká
Institute of Applied and Formal Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské nám. 25
118 00 Prague 1, Czech Republic
tel.: +42 2 21 91 42 88
fax: +42 2 21 91 43 09
e-mail:
hajicova@ufal.mff.cuni.cz
hladka@ufal.mff.cuni.cz

Contributions to TELRI newsletter, and address corrections, should be sent to:

e-mail:
hladka@ufal.mff.cuni.cz
fax: +42 2 21 91 43 09

Editorial

Wolfgang Teubert
Coordinator of TELRI

This TELRI Newsletter contains the abstracts of the presentations at the Third European Seminar: „Translation Equivalence - Theory and Practice" which will take place in Montecatini, Italy, from October 16 to 18. Like with the two preceding TELRI Seminars, it is our goal to set up a forum where industry and academia trade expertise, exchange tools and resources and prepare for the challenge of the multilingual global information society.

The synergy of 25 focal national language centers from all over Europe will give rise to new ideas and approaches for the next generation of multilingual technology: authoring tools, information retrieval and translation aids. This new generation of tools will be based on language data derived from multilingual resources: comparable and parallel corpora covering all the languages involved.

Methodologies for extracting, processing, and applying multilingual linguistic knowledge from corpora are now being developed. TELRI has undertaken a joint study on parallel texts. The results will be presented at this Seminar. Also, other speakers who are working in related relevant projects will demonstrate alternative methodologies.

We hope that the Seminar will stimulate the current multilingual NLP research and, like preceding TELRI Seminars, will lay the foundation to new joint ventures between academic institutions, language industry, and dictionary publishers all over Europe.

Topic of the issue: Multilingual technology

Multilingual Tools at the Xerox Research Centre

Jean-Pierre Chanod
Xerox Research Centre
Grenoble, France
e-mail: Chanod@grenoble.rxrc.xerox.com

The Xerox Research Centre (see http://www.rxrc.xerox.com for more information) pursues a vision of document technology where language, physical location and medium - electronic, paper or other - impose no barrier to effective use.

Our primary activity is research. Our second activity is a Program of Advanced Technology Development, to create new document services based on our own research and that of the wider Xerox community. We also participate actively in exchange programs with European partners.

Language issues cover important aspects in the production and use of documents. As such, language is a central theme of our research activities. More particularly, our Centre focuses on multilingual aspects of Natural Language Processing (NLP). Our current developments cover more than ten European languages and some non-European languages such as Arabic. Some of these developments are conducted through direct collaboration with academic institutions all over Europe.

The present articles is an introduction to our basic linguistic components and to some of their multilingual applications.

LINGUISTIC COMPONENTS

The MLTT (Multilingual Theory and Technology) team creates basic tools for linguistic analysis, e.g. morphological analysers, taggers, parsing and generation platforms. These tools are used to develop descriptions of various languages and the relation between them. They are later integrated into higher level applications, such as terminology extraction, information retrieval or translation aid. The Xerox Linguistic Development Architecture

(XeLDA) developed by the Advanced Technology Systems group incorporates the MLTT language technology.

Finite-state technology is the fundamental technology on which Xerox language R&D is based. It encompasses both work on the basic calculus and on linguistic tools, in particular in the domain of morphology and syntax.

Finite-state calculus

The basic calculus is built on a central library that implements the fundamental operations on finite-state networks. It is based on long-term Xerox research, originated at PARC in the early 1980s. The most recent development in the finite-state calculus is the introduction of the replace operator. The replacement operation is defined in a very general way, allowing replacement to be constrained by input and output contexts, as in two-level rules but without the restriction of only single-symbol replacements. Replacements can be combined with other kinds of operations, such as composition and union, to form complex expressions.

The finite-state calculus is widely used in our linguistic development, to create tokenisers, morphological analysers, noun phrase extractors, shallow parsers and other language-specific linguistic components.

Morphology

The MLTT work on morphology is based on the fundamental insight that word formation and morphological or orthographic alternation can be solved with the help of finite automata:

1. the allowed combinations of morphemes can be encoded as a finite-state network;
2. the rules that determine the form of each morpheme can be implemented as finite-state transducers;
3. the lexicon network and the rule transducers can be composed into a single automaton, a lexical transducer, that contains all the morphological information about the language including derivation, inflection, and compounding.

Lexical transducers have many advantages. They are bi-directional (the same network for both analysis and generation), fast (thousands of words per second), and compact.

We have created comprehensive morphological analysers for many languages including English, German, Dutch, French, Italian, Spanish, and

Portuguese. More recent developments include Czech, Hungarian, Polish, Russian, Scandinavian languages and Arabic.

Part-of-speech tagging

The general purpose of a part-of-speech tagger is to associate each word in a text with its morphosyntactic category (represented by a tag), as in the following example:

This+PRON is+VAUX_3SG a+DET sentence+NOUN_SG .+SENT

The process of tagging consists in three steps:

1. tokenisation: break a text into tokens
2. lexical lookup: provide all potential tags for each token
3. disambiguation: assign to each token a single tag

Each step is performed by an application program which uses language specific data:

- The tokenisation step uses a finite-state transducer to insert token boundaries around simple words (or multi-word expressions), punctuation, numbers, etc.
- Lexical lookup requires a morphological analyser to associate each token with one or more readings. Unknown words are handled by a guesser which provides potential part-of-speech categories based on affix patterns.
- Disambiguation is done with statistical methods (Hidden Markov Model), although we also experiment with fully rule-based methods.

Incremental finite-state parsing

Finite State Parsing is an extension of finite state technology to the level of phrases and sentences.

Our work concentrates on shallow parsing of unrestricted texts. We compute syntactic structures, without fully analysing linguistic phenomena that require deep semantic or pragmatic knowledge. For instance, PP-attachment, co-ordinated or elliptic structures are not always fully analysed. The annotation scheme remains underspecified with respect to yet unresolved issues. On the other hand, such phenomena do not cause parse failures, even on complex sentences.

Syntactic information is added at the sentence level in an incremental way, depending on the contextual information available at a given stage. The implementation relies on a sequence of networks built with the replace operator. The current system has been implemented for French and is being expanded to new languages.

The parsing process is incremental in the sense that the linguistic description attached to a given transducer in the sequence relies on the preceding sequence of transducers, covers only some occurrences of a given linguistic phenomenon and can be revised at a later stage.

The parser output can be used for further processing such as extraction of dependency relations over unrestricted corpora. In tests on French corpora (technical manuals, newspaper), precision is around 90-97% for subjects (84-88% for objects) and recall around 86-92% for subjects (80-90% for objects).

The LFG PARGRAM project

The LFG PARGRAM project is a collaborative effort involving researchers from Xerox PARC in Palo Alto, the Xerox Research Centre in Grenoble, France, and the University of Stuttgart in Stuttgart, Germany. The aim of the project is to produce wide coverage LFG grammars for English, French, and German which are written collaboratively, based on a common set of linguistic principles and with a commonly-agreed-upon set of grammatical features.

The grammarians use a new platform, the Xerox Linguistic Environment, which is still under development; a unification-based generator is also under development.

The grammars consist of phrase-structure rules and abbreviatory rule macros; LFG allows the right-hand side of phrase structure rules to consist of regular expressions (including the Kleene Star notation) and arbitrary Boolean combinations of regular predicates, so the rules in the grammar actually abbreviate a large set of rules written in a more conventional framework. The lexicons used by the sites consist of entries for stems, template definitions, and lexical rules. The Xerox Linguistic Environment allows for an interface to an external finite-state morphological analyser, and so the lexicons include entries for the information about morphological inflection supplied by the analyser.

APPLICATIONS

LOCOLEX: a Machine Aided Comprehension Dictionary

LOCOLEX is an on-line bilingual comprehension dictionary which aids the understanding of electronic documents written in a foreign language. It displays only the appropriate part of a dictionary entry when a user clicks on a word in a given context. The system disambiguates parts of speech and recognises multiword expressions such as compounds (e.g. *heart attack*), phrasal verbs (e.g. *to nit pick*), idiomatic expressions (e.g. *to take the bull by the horns*) and proverbs (e.g. *birds of a feather flock together*). In such cases LOCOLEX displays the translation of the whole phrase and not the translation of the word the user has clicked on.

For instance, someone may use a French/English dictionary to understand the following text written in French:

> *Lorsqu'on évoque devant les* ***cadres*** *la séparation négociée, les rumeurs fantaisistes vont apparemment toujours bon* ***train***.

When the user clicks on the word *cadres*, LOCOLEX identifies its POS and base form. It then displays the corresponding entry, here the noun *cadre*, with its different sense indicators and associated translations. In this particular context, the verb reading of *cadres* is ignored by LOCOLEX. Actually, in order to make the entry easier to use, only essential elements are displayed:

cadre I: nm
1: *[constr,art] (of a picture, a window) frame
2: *(scenery) setting
3: *(milieu) surroundings
4: *(structure, context) framework
5: *(employee) executive
6: *(of a bike, motorcycle) frame

The word *train* in the same example above is part of a verbal multiword expression *aller bon train*. In our example, the expression is inflected and two adverbs have been stuck in between the head verb and its complement. Still LOCOLEX retrieves only the equivalent expression in English *to be flying around* and not the entire entry for *train*.

train I: nm
5 : * [rumeurs] aller bon train : to be flying round

LOCOLEX uses an SGML-tagged bilingual dictionary (the Oxford-Hachette French English Dictionary). To adapt this dictionary to LOCOLEX required the following:

- Revision of an SGML-tagged Dictionary to build a disambiguated active dictionary (DAD);

- Rewriting multi-word expressions as regular expressions using a special grammar;
- Building a finite state machine which compactly associates index numbers with dictionary entries.

The lookup process itself may be represented as follows:

- split the sentence string into words (tokenisation);
- normalise each word to a standard form by changing cases and considering spelling variants;
- identify all possible morpho-syntatic usages (base form and morpho-syntactic tags) for each word in the sentence;
- disambiguate the POS;
- find relevant entries (including possible homographs or compounds) in the dictionary for the lexical form(s) chosen by the POS disambiguator;
- use the result of the morphological analysis and disambiguation to eliminate irrelevant sections;
- process the regular expressions to see if they match the word's actual context in order to identify special or idiomatic usages;
- display to the user only the most appropriate translation based on the part of speech and surrounding context.

Besides being an effective tool for understanding, LOCOLEX could also be useful in the framework of language learning. LOCOLEX also points out that existing on-line dictionaries, even when organised like a database rather than a set of type-setting instructions, are not necessarily suitable for NLP-applications. By adding grammar rules to the dictionary in order to describe the possible variations of multiword expressions we add a dynamic feature to this dictionary. SGML functions no longer point to text but to programs.

Multilingual Information Retrieval

Many of the linguistic tools being developed at our Centre are being used in applied research into multilingual information retrieval. Multilingual information retrieval allows the interrogation of texts written in a target language B by users asking questions in source language A.

In order to perform this retrieval, the following linguistic processing steps are performed on the documents and the query:

- Automatically recognise language of the text.
- Perform the morphological analysis of the text using Xerox finite state analysers.

- Part of speech tag the words in the text using the preceding morphological analysis and the probability of finding part-of-speech tag paths in the text.
- Lemmatise, i.e. normalise or reduce to dictionary entry form, the words in the text using the part of speech tags.

This morphological analysis, tagging, and subsequent lemmatisation of analysed words has proved to be a useful improvement for information retrieval as any information-retrieval specific stemming. To process a given query, an intermediate form of the query must be generated which he normalised language of the query to the indexed text of the documents. This intermediate form can be constructed by replacing each word with target language words through an on-line bilingual dictionary. The intermediate query, which is in the same language as the target documents, is passed along to a traditional information retrieval system, such as SMART[1]. This simple word-based method is the first approach we have been testing. Initial runs indicate that incorporating multi-word expression matching can significantly improve results. The multi-word expressions most interesting for information retrieval are terminological expressions, which most often appear as noun phrases in English.

Callimaque: a collaborative project for virtual libraries

Digital libraries represent a new way of accessing information distributed all over the world, via the use of a computer connected to the Internet network. Whereas a physical library deals primarily with physical data, a digital library deals with electronic documents such as texts, pictures, sounds and video.

We expect more from a digital library than only the possibility of browsing its documents. A digital library front-end should provide users with a set of tools for querying and retrieving information, as well as annotating pages of a document, defining hyper-links between pages or helping to understand multilingual documents.

Callimaque is one of our projects dealing with such new functionalities for digital libraries. More precisely, Callimaque is a collaborative project between the Xerox Research Centre and research/academic institutions of the Grenoble area (IMAG, INRIA, CICG). The goal is to build a virtual library that reconstructs the early history of information technology in France. The project is based on a similar project, the Class project, which was started by the University of Cornell several years ago under the leadership of Stuart

---

[1] This software is available for research purposes at ftp://ftp.cs.cornell.edu/pub/smart.

Lynn to preserve brittling old books. The Class project runs over conventional networks and all scanned material is in English.

The Callimaque project includes the following steps:

- Scanning and indexing around 1000 technical reports and 2000 theses written at the University of Grenoble, using Xerox XDOD, a system integrated with a scanner, a PC, a high-speed printer, software for dequeueing, indexing, storing, etc. Numerised documents can be reworked page by page and even restructured at the user's convenience. 30 Gbytes of memory are needed to store the images. Abstracts are OCRed to permit textual search.
- Documents are recorded on a relational database on a UNIX server. A number of identifiers (title, author, reference number, abstract, etc.) are associated with each document to facilitate the search
- Multilingual terminology derived from multilingual abstracts allows the system to process non-French queries.
- With a view to making these documents widely accessible, Xerox has developed software which authorises access to this database by any client using the http protocol used by the World Wide Web. The base is thus accessible via any PC, Macintosh, UNIX station or even from a simple ASCII terminal (The web address is http://callimaque.grenet.fr).
- Print on demand facilities connected to the network allow the users to make copies of the scanned material. This connection will subsequently develop towards a high output ATM network.

SELECTED REFERENCES

Salah Aït-Mokhtar, Jean-Pierre Chanod, "Incremental finite-state parsing", in *Proceedings of Applied Natural Language Processing 1997*, Washington, DC. April 97

Salah Aït-Mokhtar, Jean-Pierre Chanod, "Subject and Object Dependency Extraction Using Finite-State Transducers", *ACL workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. 1997, Madrid

D. Bauer, F. Segond, A. Zaenen. "LOCOLEX: the translation rolls off your tongue." in *Proceedings of the ACH-ALLC conference*, Santa Barbara, pp. 6-8, 1995.

Jean-Pierre Chanod, Pasi Tapanainen. "Tagging French – comparing a statistical and a constraint-based method" in *Seventh Conference of the European Chapter of the ACL*. Dublin, 1995.

Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Press, Boston, 1994.

Gregory Grefenstette, Ulrich Heid and Thierry Fontenelle. "The DECIDE project: Multilingual Collocation Extraction." *Seventh Euralex International Congress*, University of Gothenburg, Sweden, Aug 13-18, 1996.

Barbora Hladka and Jan Hajic. "Probabilistic and Rule-based Tagger of an Inflective Language". In *Proceedings of Applied Natural Language Processing 1997* Washington, DC. April 97

Ronald M. Kaplan, Martin Kay. "Regular Models of Phonological Rule Systems". *Computational Linguistics*, 20:3 331-378, 1994.

Kaplan, Ronald M. and Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*. The MIT Press, Cambridge, MA, pages 173–281.

Kaplan, Ronald M. and John T. Maxwell. 1996. *LFG grammar writer's workbench*. Technical report, Xerox PARC.

Lauri Karttunen. "Constructing Lexical Transducers". In *Proceedings of the 15th International Conference on Computational Linguistics*, Coling, Kyoto, Japan, 1994.

Lauri Karttunen. "The Replace Operator. *In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, ACL-95} 16-23, Boston, 1995.

Kimmo Koskenniemi. "A General Computational Model for Word-Form Recognition and Production. Department of General Linguistics". University of Helsinki. 1983

Julian Kupiec and Mike Wilkens. *The dds tagger guide version 1.1*. Technical report, Xerox Palo Alto Research Center, 1994.

Maxwell, III, John T. and Ronald M. Kaplan. 1991. A method for disjunctive constraint satisfaction. In Masaru Tomita, editor, *Current Issues in Parsing Technology*. Kluwer Academic Publishers, Dordrecht, pages 173–190.

John Nerbonne, Lauri Karttunen, Elena Paskaleva, Gabor Proszeky and Tiit Roosmaa. "Reading more into Foreign Languages". In *Proceedings of Applied Natural Language Processing* 1997, Washington, DC. April 97

F. Segond and P. Tapanainen. *Using a finite-state based formalism to identify and generate multiword expressions*. Technical Report MLTT-019, Xerox Research Centre, Grenoble, 1995.

A Step to Real Multilinguality (EuroWordNet and Czech Wordnet)

Karel Pala
Faculty of Informatics,
Masaryk University,
Brno, Czech Republic
e-mail: pala@fi.muni.cz

1. WHAT IS WORDNET AND EUROWORDNET?

WordNet is a database of English word meanings with basic semantic relations between them, such as synonymy, hyponymy (between expressions denoting specific and more general concepts), meronymy relations (between expressions denoting relations of parts and wholes), causal and entailment relations etc. By means of these relations all meanings can be interconnected, constituting a huge network or wordnet. Such a wordnet can be used for making various semantic inferences about the meanings of words (e.g. what words can name diseases), for finding alternative expressions or wordings, or for simply expanding words to sets of semantically related or close words in information retrieval. This approach has been developed in Princeton by G.A.Miller and his colleagues (Miller et al. 1991) and its last version is known as WordNet 1.5.

EuroWordNet is then conceived as a generic multilingual semantic database, which is the first in its kind. At present it contains the basic semantic information for Dutch, Italian, Spanish and English, while each of these resources is linked to a shared inter-lingua. This database can be directly used for semantic information-retrieval in each of these languages but also for multi-lingual retrieval across these languages. The next step is to extend EuroWordnet with a French and German wordnets so that all major European languages are covered. The basic mono-lingual databases for German and French are already being produced with national and private funding. Finally, two Eastern-Middle European languages - Czech and Estonian will be also involved in producing wordnets for them - so typologically different languages will be included as well and a standard for multilingual semantic resources for a variety of language-types will be produced.

In EuroWordNet (EWN-1) the wordnet for each language is structured along the same lines as in the Princeton WordNet 1.5 (Miller et al. 1991), in such a way that they contain synsets (set of synonymous word meanings) and basic semantic relations between these synsets. In addition each synset has an equivalence relation to a so-called Inter-Lingual-Index, mainly based on the synsets of WordNet 1.5. Via the Inter-Lingual-Index all synsets are interconnected, thus constituting a flexible and powerful multi-lingual system.

The preparation of EuroWordNet in the mentioned framework is now being designed within the EuroWordNet-2 EC Project whose co-ordinator is Piek Vossen from the Amsterdam University.

The usefulness of EuroWordNet lexical database is obvious - it will represent a resource which is essential for providing non-expert users access to the multilingual and multi-cultural European information society. Obviously, EuroWordNet is restricted to a few European languages and therefore only partially addresses the multi-linguality problem.

Furthermore, semantic networks give information about the lexicalization patterns of a language, the conceptual density of the vocabulary areas and the semantic distinctions that play a role (i.e. which meanings and which relations play a role in different semantic fields). Internet browsers are just one example of the relevance of multilingual semantic information about words for applications in the area of information retrieval. Other applications that can directly benefit from multilingual semantic resources are:

- information-acquisition tools,
- authoring-tools,
- language-learning tools,
- translation-tools,
- summarises.

A prototypical area for the application of such resources are Internet search engines, which are already well established in the information market. Although the number of users and usage of these services is increasing exponentially, the potential of the quality of results is far from being fully exploited. This holds especially for the areas of search term expansion and multilinguality. Queries are typically restricted to the enumeration (and logical combination) of mere keywords, which do not provide information about terms related to the keywords. In the present systems, a search for „health" will not disclose documents, that use clearly related terms such as „disease", „disorder", „stress", „deafness" and „headache", unless the documents also include the term „health" itself. In addition, the multilingual nature of the information society is not reflected by these engines in that they do not offer means to simultaneous access to documents in the variety of different languages, to which they have, in principle, access.

One of the main goals of EuroWordNet-2 is to include newly developed wordnets of the two Eastern-Middle European languages, particularly Czech and Estonian. The integration of these upcoming national wordnets into the EuroWordNet framework will ensure maximum compatibility between the wordnets for the individual languages and will allow true multilinguality by linking the additional resources to the shared inter-lingual database of EuroWordNet. The extension will also strengthen the role of EuroWordNet's technology and data format as a de facto standard for the representation of lexical semantic data for Europe's information society. Such a standard will not only allow for future incorporation of further languages, but also provide a unique interface for software developers in the information industry to lexical semantic data. In a longer run the wordnets will become the backbone of any semantic database of the future and will open up a whole range of new applications and services in Europe at a trans-national and trans-cultural level.

To provide non-expert users flexible access to the information society it is crucial to develop tools that can expand their general and common words in a specific language to any possible variant or term in any other language.

The user should be able to get around the choice of words in a document or the choice of key words by matching meanings rather than words. Such tools depend on the availability of generic resources with semantic information on words in each of the languages, preferably with cross-linguistic links.

2. PREPARATION OF CZECH WORDNET

The development of Czech WordNet will go along the lines outlined above. The main task is to:

a) Definition of a common set of Base-Concepts for Czech: this is a set of meanings that play a key-role in the individual wordnets. Estimated size = 1,000 synsets: 700 nominal synsets, 300 verbal synsets.

b) To encode the language-internal relations and the equivalence relations around the Base-Concepts for Czech. This should result in a Czech core wordnet of at most 10,000 synsets: 7,000 nouns and 3,000 verbs.

c) To encode the language-internal relations and equivalence relations for adjectives in Czech (and, of course, to establish links to English, Dutch, Italian, Spanish, German, French and Estonian wordnets).

d) To include Czech Base Concepts into the Inter-Lingual-Index and in this way to integrate it into EuroWordNet as its part.

The starting point for building the basic set of Czech synsets are the following resources:

i) Dictionary of Czech Synonyms (Pala, Všianský, 1994) which exists both in printed and electronic form and contains about 20 000 headwords,

ii) newly developed Electronic English-Czech and Czech-English dictionary (Ševeček, 1997) containing at present approximately 25 000 headwords,

iii)list of Czech verbs with their verb frames comprising now about 12 000 items,

iv)Czech morphological analyser and lemmatiser (Ševeček, 1996) able to retrieve the complete inventory of Czech word forms.

2.1 Techniques and/or approaches used

The scope of Czech WordNet lexical database is limited to the basic semantic relations that are well-understood - i.e. to the relations between synonyms, hyponyms, hypernyms, meronyms and holonyms plus causal relations and also verb frames.

Establishing these relations within the selected collection of Czech lexical units should be done in part semi-automatically and in part manually.

The selection of the set of Base Concepts will follow the corresponding sets in English and other languages within EuroWordNet.

The basic techniques will mainly rely on semi-automatic extracting data from above mentioned electronic resources (machine readable dictionaries) and also on using Novell toolkit.

We assume that Czech wordnet will be necessary to interconnect with the mentioned lemmatiser: in practice we have to expect that queries will have to undergo morphological analysis - this is in Czech - a highly inflected language - sine qua non for any realistic processing.

REFERENCES

Miller, G.A., et al, Five Papers on Wordnet, Princeton, 1991.

Pala, K., Všianský, J., Dictionary of Czech Synonyms (Slovník českých synonym), Lidové Noviny, Praha, 1994.

Ševeček, P., Electronic English-Czech Dictionary, Langea, Brno, 1997.

Ševeček, P., Morphological Analyser and Lemmatiser for Czech, program in C (for DOS, Unix and Macintosh platforms), Brno, 1996.

Abstracts of the papers submitted for the 3rd TELRI seminar

TRANSLATION EQUIVALENCE - THEORY AND PRACTICE

Investigating Form and Meaning Using Parallel Corpora

Michael Barlow
Department of Linguistics,
Rice University,
Texas, USA
e-mail: barlow@rice.edu
home page: www.ruf.rice.edu/~barlow

This paper provides a brief overview of some practical and theoretical issues related to parallel corpora (i.e., texts that are translations). After very brief description of a parallel concordancer, ParaConc, we will examine the potential of such a program in conjunction with monolingual text analysis programs to provide insights into the form and function of languages.

Taking a language to consist of form-meaning links, what we have in parallel corpora are two sets of form-meaning linkings, one for each language. And since the two texts are translations, the meaning part–the description of an event–can be assumed to be approximately the same in both texts. Thus we are able to see how two different languages encode equivalent meanings. The art of translation is undeniably complex, involving many different kinds of processes, and there are known problems associated with the use of translation texts, but we can fruitfully examine three main aspects of translation, namely, language particular encodings of

(i) event structure
(ii) discourse structure
(iii) lexis.

Each of these areas of form-meaning mapping can be profitably analysed using parallel corpora to yield results of interest to linguists, lexicographers, translators and language teachers. In this talk I will concentrate on the use of parallel corpora to investigate language- particular preferences with respect to the structuring of the conceptual domain in terms of metaphor and image schemas. We will see, for example, how the up-down image schema (representing the vertical dimension) is exploited to markedly different degrees in the structuring of English and French.

Performative Verbs in Plato's Republic

(with respect to their Czech-English equivalents in the text and their elaboration in the dictionaries)

Renata Blatná
Faculty of Philosophy,
Charles University,
Prague, Czech Republic
e-mail: Renata.Blatna@ff.cuni.cz

The topic of this contribution was chosen in accordance with the nature of the text studied - the dialogue form: the participants react to each other's statements, questions etc. Therefore, rather great number of verbal forms in the 1st person, present indicative appear in the text, such as *I tell you, I ask you, I agree, I admit* x *říkám, tvrdím, souhlasím* etc. The performative verbs were defined by J. L. Austin in his book „How to do thing with words" (1962) and then by J. R. Searle in his work „Speech acts" (1969). The theory of speech acts was summarized in the work „Pragmatics" by S. Levinson (1983). Levinson criticized Searle's typology of speech acts and stated that „the ‚fundamental part' part of human communication is carried out... by specific classes of communicative intention" (p. 241). According to Levinson there are three basic sentence-types, i. e. interrogative, imperative and declarative and they seem to be the universal of most of the languages. These three sentence-types may contain the performative phrases or prefixes, e. g. *I request you to*, i. e. explicit performative verbs. These sentence-types were taken as the basis for our analysis.

The declarative sentences are represented in the greatest number of occurrences. Most frequent performative forms in Czech are the following:

tvrdím / netvrdím	16x
říkám	10x
souhlasím	10x

shoduji se s tebou	6x
jsem zajedno	1x

Most frequent English equivalents: *I affirm, I say, I agree, I admit, I tell you*

As an example we can take the text equivalents of the Czech performative „*(já) ne/tvrdím*":
I affirm that 3x, I say 4x, I concur 1x, I (don't) mean 2x, I am trying to say 1x, I am ready to admit 1x, we say 1x, I will say 1x.

Other contexts:

1) Pak nebudeme múzicky vzdělaní, *tvrdím* při bozích, ani my...
Then, by heaven, *am I not right in saying* that by the same token we shall never be true musicians, either

2) ... pokud ty tady netvrdíš něco jiného. - *Netvrdím*, řekl.
unless you have something different to say." - „No, *nothing*," said he;

3) tu *tvrdím*, nemohl by k tomuto poznání nikdy dojit
I would never say that he really learns

One of the text variants which appears as an equivalent of other verbs, is, of course, the auxiliary verb, i. e. *I do*. According to the Czech-English dictionary by I. Poldauf the equivalents of *tvrdit* are: *to insist, to claim, to assert, to affirm, to allege, to aver, to predicate, to vindicate, to submit, to argue, to contend, to maintain, to warant*. It is quite interesting that among these equivalents the most common text equivalent *to say* does not occur.

In the similar way within the contribution the interrogative and imperative sentences will be analysed.

Translation Equivalents: Where Neither A Dictionary Nor A Corpus Helps

Eva Hajičova, Zdeněk Kirschner
Charles University, Prague
e-mail:{hajicova, kirschner}@ufal.mff.cuni.cz

1. Adding new words to the lexical stock of natural language is an endless process. Therefore, every lexicon is an open list, even if based on corpora of

hundreds of millions of word occurrences. For purposes of multilingual applications one needs to think of „fail-soft" measures to cover the text as a whole, without blanks substituted for the unknown words.

One possible solution is to study productive word formation processes as a basis for an automatic „transduction" of the given unknown lexical unit of the source language.

2. Such a transducing device was developed by Zd. Kirschner within the project of English-to-Czech machine translation in the eighties. It was based on the observation that most newly coined Czech words in the domain of technology and science are taken over from English, as loans from Latin and Greek with slight (and mostly regular) modifications as for endings and orthography. Based on this observation, a set of about 60 rules was formulated to cover the most productive modifications.

The first step consists in the interpretation of the unrecognized words according to their typical and (mostly) productive suffixes (the inflectional endings being detached and dictionary forms reconstrcuted by morphemic analysis in the preceding steps), and to assign them the POS and semantic information. Thus e.g. words ending in *-er, -or, -graph, -ode* and some others are interpreted as nouns, concrete, denoting actors/instruments (e.g. *adapter, detector, cyclograph, cathode*); words ending in *-ce, -cy, -ess, -tude* are supposed to be nouns, abstract, properties and forming a regular adjective in Czech (*equivalence - ekvivalence, ekvivalentní; tendency - tendence, tendenční; absurdness - absurdnost, absurdní; altitude - altituda, altitudní*); the same characteristics are assigned to unknown words ending in *-ity, -ism, -ship, -hood, -thm*, except for the morphemic information on the formation of adjectives (*selectivity - selektivita,* *selektivitní, *isomorphism - izomorfismus,* *izomorfizní; *dicatorship - diktátorství*, etc.); the endings *-fy, -ate, -ise(-ize), -duce* indicate verbs that can be both transitive and intransitive, of causative and (semi)terminological character, yet not allowed to form adjectives of the purposive character (*calcify - kalcifikovat; alternate - alternovat; formalize - formalizovat, induce - indukovat*). A number of adjectival endings was covered by the rules as well, e.g. *-ary, -al, -rse, -ive, -ous, -ic, -ble, -less, -anar, -lear, -near, -olar, -ular* (*evolutionary - evoluční; global - globální; disperse - disperzní*). The transducing device covers about 50 classes of nouns, 13 classes of adjectives and 4 classes of verbs.

In the next step, the English suffixes are replaced by the Czech ones, and, finally, the word bases are scanned for spelling configurations to be transformed or adapted to Czech orthography. Thus (as the above examples illustrate), e.g., *ph* is replaced by *f*, *th* by *t*, *c* preceding *a, l, o, r, t, u* by *k*, *s* preceded by *a, e, i, n, o, r, y* and followed by *a, e, i, o* is replaced by *z*, etc.

To give some more examples, *photolithographic* is translated as *fotolitografický*, *cyclotron* is translated as *cyklotron*, *operational* as *operační*, etc. This is not to say that the transduction always results in an existing Czech word, but in most cases it does, and in most of the remaining cases the transduction leads at least to a satisfactory classification of the word as for its POS and its morphemic properties, so that the specialist of the domain gets a reasonable picture of the structure of the whole sentence (e.g. the English word *amplifier* would be translated as *amplifikátor* rather than as the correct *zesilovač*, but this transduction would not lead to a misunderstanding on the side of the reader).

Such a transduction device, of course, must be based on a careful empirical analysis of word formation in the given pair of languages; otherwise, the process may result in unpleasant misinterpretations. Thus, in one of the beginning phases of our experiments, the transduction procedure checked first on texts from the domain of electronics, was applied to a more general domain, where a source text included the collocation *international conference*. Since one of the rules rewrites the ending *-ational* to the Czech ending *-ační*, the resulting translation was *internační konference* rather than the correct *internacionální konference* (*inter-national*); however, the adjective *internační* does exist in Czech with the meaning ‚internment': an interment conference (especially under the totalitarian regime) is far from an international conference.

To estimate the scope of coverage of the transducing procedure formulated within the mentioned MT project, we have scanned the inverse dictionary of English and counted how many words would be correctly treated by the transducing device. The set of about 60 rules covers about 20000 lexical entries from the dictionary.

4. These good results have encouraged us to try and test this fail-soft measure in the Czech-to-Russian MT system developed by our research team (Bémová and Kuboň 1990). The initial expectation was that with languages that are closely related to each other the idea of a transduction dictionary could be applied in an even larger range. A contrastive analysis of Czech and Russian has shown that many items actually can be translated in the above illustrated algorithimc way. A large class of Czech words can be translated into Russian by a mere transcription (at least in the nominative or nom./acc. case), cf. e.g. *elektroskop - elektroskop, expozimetr - ekspozimetr, cyklograf - ciklograf, agregát - agregat, demontáž - demontaž*. Another group contains words the form of which must (also) be modified by a regular procedure, cf. e.g. the derivation

suffixes and inflectional endings in *formalismus - formalizm*, *linearizace - linearizacija*, *extrakce - ekstrakcija*, *tendence - tendencija*, *stenogram - stenogramma*, *homeostaze - gomeostazis*, *báze - bazis*, *hypotaxe - gipotaksis*, *galium - galij*, *selektivita - selektivnost'*, *helium - gelij*, *specialista - specialist*. The third group includes semantically uniform and productive classes of words, as e.g. deverbative nouns (*-ání -> -anie*), nouns denoting a property (*-ost -> -ost'*), nouns with the meaning of a certain place or space (*-ště -> -šče*), nouns with a meaning with a feature of property (*-tví -> -tvo*).

However, in the course of a long-term development of these languages, the semantic shifts of the word bases prevent the possibility of translation of these types only by means of the word-formation correspondences of the transductive dictionary. This point can be illustrated on the example of deverbative nouns in *-ání*, *-ení*: *projektování = projektirovanie* (designing), *referování = referirovanie* (refereeing), but *simulování = imitacija* or *modelirovanie* (simulation or modelling) rather than *simulacija*. In several cases it is possible to apply the regularity of sound changes between Czech and Russian: the Cz. prefix *pře-* can be transduced as *pere-* (*přejmenování = pereimenovanie*), but we also face such cases as *přetečení* (= *perepolnenie*, overflow), *přepínání* (*pereključenie*), which cannot be translated in such a mechanical way. In such cases, the transducing dictionary cannot do more than specify the word class or the gender, that is the information to be used in the syntactic analysis of the source language, but in the Russian output the word has to be marked as „not found in the dictionary".

The above remarks are just an illustration of the possibilities and limitations of an application of a transducing procedure to a pair of closely related languages: our experience indicates that with closely related languages, there is a bigger danger of „coining" false equivalents than with a pair of languages that belong to different families but share the tendency to coin new words from the same (Latin or Greek) basis.

Nevertheless, we hope to have illustrated that even with large-scale multilingual corpora one has to look around for some fail-soft measures that take care of the outcome of the dynamic processes of the formation of neologisms. One of such measures has been described in this contribution.

References:

Bémová A. and V. Kuboň (1990), Czech-to-Russian Transducing Dictionary. In: COLING-90, Peprs presented to the 13th Int. Conference on Computational Linguistics, Helsinki, 314-316.

Hajičová E. and Z. Kirschner (1987), Fail-Soft („Emergency") Measures in a Production-Oriented Mt System. In: Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics, Copenhagen, 104-108.

Word Sense Disambiguation with Multi-Lingual Corpora

Nancy Ide
Vassar College, USA
e-mail: ide@cs.vassar.edu

Word Sense Disambiguation (WSD) is one of the foremost problems facing research in natural language processing today. Polysemous words present obstacles in areas as diverse as machine translation, document retrieval, and speech synthesis. Recent work on WSD suggests that aligned parallel corpora offer a ready-made solution to sense disambiguation, since the translation of different senses of a polysemous word often differs. For example, the word „sentence" in English is translated in French as „phrase" in its sense as a grammatical construct, and as „peine" in its sense as a prison term. To disambiguate an occurrence of the word „sentence" in an aligned English-French corpus, then, one need only consult the translation in the French to see which translation is used. However, this disambiguation method is only partially reliable, for several reasons. First, in many cases the ambiguity is preserved in the translation (e.g., „interest" in English is „interet" in French regardless of its sense). Second, translation is not always word-for-word, and semantic mappings may vary with subtleties of use, etc.

So far, all work using parallel corpora for WSD has involved alignment between only two languages. However, the availability of parallel corpora in multiple languages (e.g., the Republic of Plato and Orwell's „1984" in several languages) offers new potential for exploiting this resource in WSD work. Such corpora provide even more potential because they involve translations in languages from different linguistic families, in which sense ambiguity is less likely to be preserved, and where it is more likely that at least one parallel text could provide the information required for disambiguation. The potential for the use of such corpora for WSD needs to be systematically explored, in order to determine how many and which kinds of languages are required for effective WSD and which kinds of information is necessary to extract from the parallel translation; to identify potential problem areas; to develop appropriate methodologies; etc. This paper is intended to assess the potential of multiple parallel translations for WSD, and provide some principles and methods based on the results.

CAT-Tools - Computer Aided Translation Tools

Iris Jahnke
Trados (Schweiz) AG,
Bern, Switzerland
e-mail: iris@trados.ch
http://www.trados.ch

Whereas in the past, automation of the professional translation process was mostly connected to the use of machine translation (MT), this has significantly changed in the last few years. Today, the keywords for professional translators are *computer aided translation tools* (*CAT-Tools*) and, notably a key-component: the *translation memory*. The general idea of a translation memory is very simple: All translations made by a translator are stored in a database and are then in case of re-translations immediately retrievable.

Modern CAT-Tools, in most cases an integration of several functionalities into one „workbench", are gaining more and more ground as a standard tool in the hand of professional translators. Except for literary translations or generally idiosyncratic text types, the use of CAT-Tools has been extended to almost every type of translation work. This includes political, administrative, technical, advertising, biographical, and other text types.

Nowadays companies are faced with a rapidly growing volume of documentation that needs to be produced with ever shorter production cycles while still maintaining the high quality standards expected by its international clients.

This is one of the many reasons why the Trados CAT-Tools are used by companies and institutions like the European Commission and Microsoft. These tools consisting of a terminology database and a translation memory system, make translation work much more efficient. Clients using the Translator's Workbench, Trados translation memory system, speak of timesavings between 30%-50% on text with a certain percentage of repetitiveness and a higher quality standard.

But lets have a closer look at the following products on the basis of a translation project: „computer manual - English > German"

- Terminology database: MultiTerm
- Translation memory system: Translator's Workbench

And what does the future bring?

Terminology search on the Internet/Intranet possible today - with MultiTerm Web Interface.

This will be shown on the basis of a worldwideweb on-line search on the database of the European Parliament „Euterpe" or the database of the Credit Suisse.

TRADOS, founded in 1984, based in Stuttgart (Germany), develops and markets tools for professional translators, providing a full range of products and services in this field. Today, with a network of sales and support offices throughout Europe and the US, TRADOS is considered to be one of the leading tools vendors in this market.

Probabilistic Tagging in a Multi-Lingual Environment: Making an English Tagger understand Romanian

Oliver Mason
School of English,
University of Birmingham,
Birmingham, Great Britain
e-mail: oliver@clg9.bham.ac.uk

Dan Tufis
Center for Advanced Research in Machine learning, NLP
and Cognitive Modelling,
Bucharest, Romania
e-mail: tufis@valhalla.racai.ro

This talk describes the process of adapting a parts-of-speech tagger originally developed for English to work language independent. It is shown that a probabilistic tagging approach works well if the language specific information can be separated from the processing engine. An evaluation has been done on Rumanian data which showed encouraging results. With only about 200K words of training data a rate of 97.5% correct tag assignments could be achieved.

Technical-Scientific Translation: An Industrial Necessity

Gianluca Mattioli
Consortium for the Training & Development of SMEs,
Bologna, Italy
e-mail: mattioli@cofimp.it

The paper generally deals with the problem of translating technical-scientific documents within the area of the Small and Medium Industries in Italy exporting abroad. To master the art of translating, not only from the simple practical point of view, is a necessity felt ever more deeply by the industrial world, where foreign trading is becoming an essential part of the economy. The modern post-industrial society lives on communication, and documentation constitutes the most important information vehicle; a document able to concretely inform the specialised reader in fact transforms the latter in a confident and knowledgeable user.

In the case of technical document translation, the fast evolution of specialist languages makes dictionaries obsolescent and terminologically inadequate - and these dictionaries are the traditional sources of reference, often still the translator's only working tools. However, in spite of the uncertainties of dictionaries, the scarcity of alternative reliable sources makes it possible for them to be regarded as gospel, with the imaginable poor results.

What makes a technical-scientific document hard to translate is mostly the lack of sure definitions and reliable terminological sources and references. Furthermore, terminological work presents the typical difficulties of a strongly comparative and relational activity, and terminological analysis constitutes the initial and primary part of a technical translator. The quality of the terminology employed in a technical document is determined by its definition level and influences the degree of uniformity and coherence achievable, and thus the degree of text ambiguity. Often one makes indiscriminate use of jargon expressions, most of the time conforming only to realities within their own texture, exactly because used in an inappropriate or incorrect manner.

A terminological data bank can make available to users (even non-advanced ones) a tool of easy and prompt consultation, yet at the same time efficient and exhaustive, complementing the traditional references and flexible in its updating when employed in systematic translation. The small and medium-sized enterprises in Italy, and especially in Emilia Romagna, are ripe for being properly introduced to advanced systematic translation and specialised data banks, and would greatly benefit from them if educated to understand their principles and advantages. The author expresses his interest for any suggestions originating from the TELRI Seminar, and would be ready to disseminate any relevant information to the industrial world of Emilia Romagna, keeping also in mind the current legislation relating to linguistic requisites within the EEC.

It is clear that, besides competences in the foreign language (certainly not only of grammatical and lexical nature), a translator would need to possess specific skills of five broad orders:

1) encyclopaedic knowledge of the topic treated;
2) capacity to identify and manipulate concepts;
3) knowledge of textual strategies;
4) expressive capacities of writing in the target language;
5) capacity to manipulate transcultural phenomena.

To the author's knowledge, no complete didactic project exists able to articulate in a progression these five competencies and thus satisfy the current autonomous needs of Italian SMEs. Educational institutions give priority to training in foreign languages, both in the university curricula and in the translation/interpreting schools, complemented by notions on „general culture", „civilisation" of the country of origin and study of international organisations. All this is quite inadequate, if we look at the translating problems currently existing in the field, and the result basically is that apprentice translators deem knowledge in a foreign language still the decisive factor.

It is a fact that the skills for translating are not acquired solely by learning the foundations of a foreign language; solid linguistic knowledge should be complemented by a good acquaintance with the topic to be dealt with, as well as a noticeable dose of precision and creativity. To write, in fact, is still an arduous and demanding task, and acquaintance with the subject considered is fundamental to the elaboration of technical documents, because by this depends the capacity of properly transferring technical-scientific information.

Documentation is the interface between user and product or service, and should place the user in the condition to use it, not only by transmitting information, but also by combining the product's or service's functions with the user's needs and his expectations, and thus it should constitute its integrating element.

Within this perspective, the paper asserts that translating should be deemed a product which is modular and functional to that which it integrates, such that it can be considered the result of an independent activity, yet essential and inseparable part of the product and its generating process. Therefore, perfect transposition cannot be achieved outside the product to which it is associated, and each process phase should be guided by the requisites of the final product. Similarly, the operators of such transposition ought to be deemed part of an integrated group and their activity should be taken into consideration during the product planning stages as well as during the stages defining the producing cycle and life, so that the different product's versions may follow the same evolution of the original product.

The **third Chapter** discusses more at length some points excerpted from the mentioned statistical research conducted by COFIMP in April 1996: it appears evident that there is a great need for appropriate technical translating systems and expertise among the SMEs of Italy, which are constantly expanding their export markets to now include eastern European countries and Asia. Yet there is a reluctance on their part to approach the problem - since *it is a problem*, due to the SME's lack of adequate translating structures - in a serious and professional way, because thus far the SMEs have not really been informed of the actual resources currently available. COFIMP's research has shown that, however, the SMEs have a clear concept of what they would require of a „translator“, should they accept an in-house presence (be it physical or „electronic“) instead of random subcontracting their language requirements to insecure local Translating Agencies. This Chapter should be read **in conjunction** with the actual COFIMP Survey, available in hardcopy during the course of the Seminar.

Form and Sense Relations as Seen Through Parallel Corpora.

Anna Mauranen
University of Joensuu
Savonlinna School of Translation Studies
Savonlinna, Finland
e-mail: mauranen@joyl.joensuu.fi

The paper starts from the position that translated texts constitute a valuable component of any representative corpus of a natural language; the contrast between languages is seen as relevantly embodied in the practices of bilingual users. The particular focus of the paper is on the value of parallel corpora for contrastive language study in the light of a corpus of English texts and their translations into Finnish. By taking a single common lexical item as a point of departure (the lemma *think*), the paper shows that the translation equivalents in the corpus have a different profile for each of the forms of *think*. The target language equivalents provided by professional translators in real contexts can thus be seen as reflecting the sense profiles of the source language word forms. This finding throws doubt on the common practice in contrastive analysis of taking the equivalence between lemmas as the basis of comparison, and, by extension, on the usual practice of compiling bilingual dictionaries.

In addition to reflecting the source language, the juxtaposition of a source language and translations also allows insights into the target language: for example, the study reported here discovered certain delexicalised uses of a major Finnish equivalent of *thought*. A parallel corpus can thus be seen as a unique source of insights into both the languages concerned; as well as offering material for developing hypotheses for further testing with monolingual corpora, it also provides a data-driven starting-point for contrastive analysis.

Automatic Extraction of Translation Equivalents from Aligned Corpora of Legal Texts

Stoyan Mihov
Linguistic Modeling Laboratory,
Bulgarian Academy of Sciences,
Sofia, Bulgaria
e-mail: stoyan@lml.acad.bg

The paper describes the system MARK-ALISTeR for automatic alignment and search of translation equivalents in large bilingual corpora. In MARK-ALISTeR the Gale-Church algorithm is chosen as an aligning procedure for parallel texts and the Ted Dunning's method based on likelihood ratios was adopted for searching of translation equivalents. Special attention is paid to the extension of the system for searching exact translation equivalents of words and phrases. This implementation is related to BILEDITA #790 Copernicus'94 Joint Research Project where a French-Bulgarian Bilingual terminological dictionary was automatically extracted from parallel legal texts. Evaluation of the results of searching translation equivalents is presented.

TRACTOR – TELRI Research Archive of Computational Tools and Resources

Ramesh Krishnamurthy
School of Humanities, University of Birmingham
Birmingham, United Kingdom
e-mail: ramesh@cobuild.collins.co.uk
http://www-clg.bham.ac.uk

The establishment of TRACTOR was always a part of the long-term TELRI aims and was outlined in early drafts of the TELRI Workplan. In the past few months, the Archive has begun to take shape, thanks to the generosity and cooperation of TELRI partners and others, such as the Le Monde Diplomatique organization.

Two temporary full-time workers have been appointed to bring the project to completion: Ramesh Krishnamurthy (former Corpus Manager of Cobuild at Birmingham University) and Chris Kidd (a computer specialist with previous experience of large-scale projects for, among others, British Telecom and the UK Department of Health).

The list of resources offered so far includes an impressive variety of languages, text-types, and software tools. Over 20 languages are represented individually, as well as parallel texts in several languages. When properly analysed into their component units, there will probably be well over 300 individual items. Much of the data is of recent vintage, but there is a good range of historical material as well.

Data-types are primarily from written sources, including newspapers, textbooks, dictionaries, literary works, grammars, academic, scientific and technical texts, popular fiction, poetry, jurisprudence, government and military records, and so on. But there are also a few texts of transcribed speech and even some digitized sound files.

Computational tools and resources present an even greater variety: de-asciifiers, spelling correctors, morphological analyzers and generators, concordancers, compound-recognizers, foreign-word-recognizers, general corpus retrieval and analytical tools, taggers, parsers, segmenters, editors, character-combination frequency lists, etc.

Texts vary considerably in size, from millions of words of newspaper data to small texts of a few hundred words. It is difficult to estimate the final overall size of the archive, but it is likely to be well over 100 million words.

IDS-Mannheim have made available to TRACTOR their solaris3 server with 23 gigabytes of storage, and Ramesh and Chris have commenced their formidable task of transferring all the TRACTOR data and tools to that machine. However, at the same time, all the data texts are being converted to PAROLE-standard conformance, and all tools are being supplied with standardized documentation.

By the end of December 1997, all the resources will be catalogued and made available to TELRI partners and other entitled users. From now on, progress will be reported and constantly updated at the new TRACTOR Website.

Using Parallel Corpus in a Translator Training Program

Daniel Ridings
Göteborg University,
Göteborg , Sweden
e-mail: ridings@svenska.gu.se

It goes without saying that parallel texts must offer a wealth of information that can be used in a translation context. Much energy is spent on isolating translation equivalents for words from the general language and technical terms. Most of these approaches get us quite far, but rarely far enough that we can say for sure exactly what is equivalent to what below the level of a „sentence." These problems are continually supplying us academics with the raw material for further research and investigation.

In the meantime, the translators are waiting for the simple tools we promised them. This presentation will exemplify how preliminary results of research projects are lifted out of their academic context and combined with tools that are already on the market in order to offer translators assistance over and beyond what is available in paper format. This is illustrated by showing how prelimary results from the parallel text project in Gothenburg have been integrated with MULTITERM, a commercial terminological management system from TRADOS.

The main points that will be dealt with are the implications that corpus-based multilingual lexicography has on the structure of the lexical database and on how we implement some tentative results of collocational studies into a production environment using MULTITERM for students from the translator training programme at Göteborg University.

Information System RUSSIA
- the Thesaurus on Modern Life in Russia

Tatyana Yudina
Center for Information Research,
Moscow State University,
Moscow, Russia
e-mail: cir@online.ru

The bilingual Thesaurus on modern life in Russia is part of the Information System RUSSIA project. The Thesaurus is being developed as a component of the NLP-technology and serves both for full text documents' indexing and categorization and as a search instrument. It is being translated into English in order to provide for foreign specialists to retrieve the IS RUSSIA and as a tool to search the Internet sites' documents in English and to produce index and event-categorization of them in Russian. The Thesaurus on Modern Life in Russia incorporates more than 30,000 linked entries (it includes geographical part of 7,000 entries), it is created by mutual work of programmers, linguists and experts in social, political, economic sciences. 150 Mb of Russian political texts were processed in a semi-automatic mode to produce thesaurus entries. Thesaurus translation resembles with the most sophisticated ones - the Legislative Indexing Vocabulary of the US Congressional Research Service, L.C.; LegiSlate Thesaurus, United Nations Thesaurus, WestLaw Thesaurus, the EVROVOC (thesaurus of the Commission of European Communities). It is also arranged to meet the standards enforced by UNESCO to ensure its international compatibility.

The Information System RUSSIA (IS RUSSIA) is an integrated computer-based information resource for Internet access to data and documents on government and politics in the Russian Federation. The IS RUSSIA project has initially pursued the main goal to create a free computer-based library for general public access, functioning as a data archive for research and education in human sciences. Special part of the project - the NLP-technology - provides for automatic processing of large scope of data and value-added (analytical) services. This component is especially important for human studies given how large the volume of information (including full text documents) are to be processed daily to monitor and analyze the social developments. A pecial part of the project is bilingual complex that includes friendly interface and help screens, developed search tools and abridged versions of reference databases. The thesaurus-based search tools allow advanced query expansion based on the concept relationship encoded in the thesaurus. It makes the search more intelligent, efficient, rational, time- and cost-saving.

The IS RUSSIA project is being developed by a non-commercial organization - the Center for Information Research and is housed at the Scientific Computer Center of the Moscow State University. Financial support was provided by foreign charitable funds, Russian government and scientific funds: the MacArthur Foundation, USA, (1993, 1995, 1996), the Ministry of Science of Russia (1995, 1996), the Ford Foundation, USA, (1996), the Russian Fund for Fundamental Research (1997), the Russian Humanitarian Scientific Fund (1997). Two specialists working with the team have received individual grants from the Soros Foundation in 1995 and the MacArthur Foundation in 1997.

The IS RUSSIA is available on the Internet (http://terminus.srcc.msu.su) since April 1997. User access is currently limited by the hardware capabilities.

On TELRI Newsletter*

Eva Hajičová, Barbora Hladká

The main task of the working group "Newsletter" was to prepare and publish TELRI Newsletter in regular intervals (three times per year) to inform the academic community, their industrial partners and also the prospective users about the activities of individual TELRI working groups, about available resources and about methods for their processing.

---

* Contact the editors to receive the complete set of Newsletter or particular issues.

The first issue of the Newsletter was printed and distributed for the September 1995 Tihany meeting. Thanks to this issue we could meet with Trans-European Language Resources Infrastructure, i. e. with TELRI partners, working groups and planned TELRI events.

The second issue (December 1995) was prepared and ready to be sent out in December 1995 and was devoted to the first European Seminar, "Language Resources for Language Technology" conducted in Tihany, Hungary. Demonstrations of NLP systems of most different kinds were one of the most interesting part of the Seminar. No. 2 brings short descriptions of demonstrations and contributions devoted to some joint ventures.

In 1996, the issues 3 (June 1996) and 4 (October 1996) were put together, edited and printed. We introduced a new column called "topic of this issue" in No. 3. The first discussed topic was "syntactic tagging". In No. 4, we continued with discussing "syntactic tagging". The Nancy workshop was an example of joint activities between the members of the working groups. Some workshop participants´ remarks are presented on the pages of the that issue.

The contents of No. 5 (April 1997) described mainly the results of the Ljubljana workshop, which was concentrated on the work on electronic text version of the sample text, Plato´s "Republic".

The second European Seminar, "Language Applications for a Multilingual Europe" was held in Kaunas, Lithuania. Newsletter No. 6 (August 1996) was mainly focused on the descriptions of the demonstration in Kaunas. Like in No. 5 the lexicons were the core theme of the issue.

The present issue (No. 7) is the last issue of TELRI Newsletter and is devoted mainly to the third European Seminar, "Translation Equivalence - Theory and Practice", which takes place in Montecatini, Italy. The main topic of the seminar - multilingual aspects of corpora processing - is reflected in most of the contributions to this issue.

We would like to acknowledge the efforts of all contributors who made our work easier. We hope that the Newsletter has been a useful and functional link between the partners and the language engineering community.

October 1997

List of Participants[*]:

ANDERSEN Poul		*e-mail:*	m764@eurokom.ie
BECI Bahri		*e-mail:*	beci@igjl.tirana.al
BENKO Vladimír		*e-mail:*	jazybenk@savba.savba.sk
BIEN Janusz S.		*e-mail:*	jsbien@plearn.edu.pl
ČERMÁK František		*e-mail:*	frantisek.cermak@ff.cuni.cz
ERJAVEC Tomaž	*NEW!!!*	*e-mail:*	Tomaz.Erjavec@ijs.si
FISIAK Jacek		*e-mail:*	fisiak.plpuam11.bitnet
GELLERSTAM Martin		*e-mail:*	gellerstam@svenska.gu.se
HAJIČOVÁ Eva,		*e-mail:*	hajicova@ufal.mff.cuni.cz
HLADKÁ Barbora		*e-mail:*	hladka@ufal.mff.cuni.cz
JAKOPIN Primoz		*e-mail:*	primoz.jakopin@uni-lj.si
JAROŠOVÁ Alexandra		*e-mail:*	sasaj@juls.savba.sk
KRUYT Truus		*e-mail:*	kruyt@rulxho.leidenuniv.nl
LAURENT Romary		*e-mail:*	Laurent.Romary@loria.fr
LAWSON Ann	*NEW!!!*	*e-mail:*	lawson@ids-mannheim.de
MARCINKEVIČIENĖ Rūta		*e-mail:*	ruta.marcinkeviciene@vdu.lt
OIM Haldur		*e-mail:*	hoim@psych.ut.ee
PAJZS Júlia		*e-mail:*	pajzs@nytud.hu
PASKALEVA Elena		*e-mail:*	hellen@bgearn.bitnet
PENCHEV Iordan		*e-mail:*	jpen@bgearn.bitnet
SINCLAIR John M.		*e-mail:*	j.sinclair@bham.ac.uk
SPEKTORS Andrejs		*e-mail:*	aspekt@mii.lu.lv
TEUBERT Wolfgang, VOLZ Norbert		*e-mail:*	telri@ids-mannheim.de
TUFIS Dan	*NEW!!!*	*e-mail:*	tufis@valhalla.racai.ro
ZAMPOLLI Antonio		*e-mail:*	paula@icnucevm.cnuce.cnr.it

---

[*] You can see detailed addresses in Newsletter No. 2.

WHAT IS TELRI

The Concerted Action TRANS-EUROPEAN LANGUAGE RESOURCES INFRASTRUCTURE (TELRI), is a COPERNICUS project funded by the European Commission. TELRI has a duration of three years (1995-1997). It brings together 22 institutions of 17 European countries (Albania, Germany, Great Britain, Slovakia, Italy, Bulgaria, the Czech Republic, Sweden, Slovenia, Romania, Estonia, France, the Netherlands, Latvia, Lithuania, Poland and Hungary).

TELRI is setting up a permanent network of leading national language and language technology centres in the whole of Europe. It pools existing language resources, corpora, machine-readable dictionaries and lexicons, lexical databases, and generic software tools for the creation, re-use, maintanance, validation , and exploitation of linguistic data. It complements these repositories with newly created multilingual resources, offering a wide range of language data to the NLP community. TELRI is establishing a platform where research and industry meet, exchange resources and engage in product-oriented cooperation.

Links have been established with language centres elsewhere in Europe, with relevant European organizations and ventures, and with focal language institutions in other parts of the world.

FOR INFORMATION.

Inquiries about TELRI may be addressed to: **Dr. Wolfgang Teubert,** Institut für deutsche Sprache, P. O. Box: 101621, 68016 Mannheim, Germany, Phone: +49 621 1581 437, Fax: +49 621 1581 415, e-mail: telri@ids-mannheim.de

TELRI's WWW Document.

Detailed information about TELRI and its activities is available through the World Wide Web (WWW) at the following URL: **http://www.ids-mannheim.de/telri/telri.html**
Webmaster: Alena Bohmová, e-mail: webadm@smetana.ms.mff.cuni.cz
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Malostranské nám. 25, 118 00 Prague 1, Czech Republic